

# Data Driven Resource Allocation for Distributed Learning\*

Travis Dick

Carnegie Mellon University  
tdick@cs.cmu.edu

Venkata Krishna Pillutla

Carnegie Mellon University  
pillutla@cs.cmu.edu

Maria Florina Balcan

Carnegie Mellon University  
ninamf@cs.cmu.edu

Mu Li

Carnegie Mellon University  
muli@cs.cmu.edu

Colin White

Carnegie Mellon University  
crwhite@cs.cmu.edu

Alex Smola

Carnegie Mellon University  
alex@smola.org

December 16, 2015

## Abstract

In distributed machine learning, data is dispatched to multiple machines for processing. Motivated by the fact that similar data points are often belonging to the same or similar classes, and more generally, classification rules of high accuracy tend to be “locally simple but globally complex” [33], we propose data dependent dispatching that takes advantage of such structures. Our main technical contribution is to provide algorithms with provable guarantees for data-dependent dispatching, that partition the data in a way that satisfies important conditions for accurate distributed learning, including fault tolerance and balancedness. We show the effectiveness of our method over the widely used random partitioning scheme in several real world image and advertising datasets.

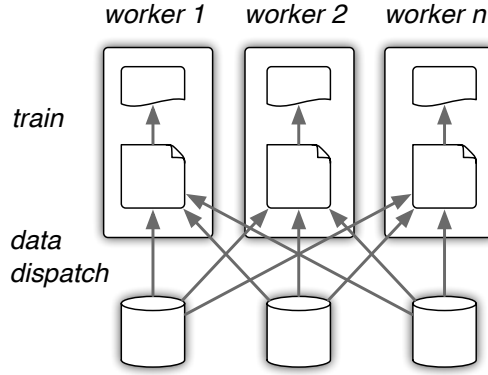
## 1 Introduction

**Motivation** Distributed computation is playing a major role in modern large-scale machine learning practice with a lot of work in this direction in the last few years [4, 5, 6, 25, 34, 35]. This tends to take two high-level forms. The first is when the data itself is collected in a distributed manner, whether from geographically-distributed experiments, distributed sensors, distributed click data, etc., and the goal is to take advantage of all this data without incurring the substantial overhead of first communicating it all to some central location. The second high-level form is where massive amounts of data are collected centrally, and for space and efficiency reasons this data must be dispatched to distributed machines in order to perform the processing needed [25, 35]. It is this latter form that we address here.

When data is dispatched to distributed machines, the simplest approach and what past (both theoretical and empirical) work has focused on is to perform the dispatching randomly [34, 35]. Random dispatching has the advantage that dispatching is easy, and also, because each machine is receiving data from the same distribution, it is rather clean to analyze it theoretically. However, since in the end the statistical serving model on all machines is essentially identical, such techniques could lead to sub-optimal results in practice in terms of the accuracy of the resulting learning rule. Motivated by the fact that in practice, similar data

---

\*This work was supported in part by NSF grants CCF-1451177, CCF-1422910, CCF-1535967, IIS-1409802, a Sloan Research Fellowship, a Microsoft Research Faculty Fellowship, a Google Research Award, Intel Research, Microsoft Research, and a National Defense Science & Engineering Graduate (NDSEG) fellowship.



**Figure 1:** Data is partitioned and dispatched into multiple workers. Each worker then trains a local model using its local data. There is no communication between workers during training.

points tend to have the same or similar classification, and more generally, classification rules of high accuracy tend to be “locally simple but globally complex” [33], we propose a new paradigm for doing *data dependent dispatching* that takes advantages of such structures.

In particular, we introduce and analyze dispatching techniques that partition a set of points in such a way that similar examples end up on the same machine/worker, while satisfying key constraints present in a real world distributed system including balancedness and fault-tolerance. Such techniques can then be used within a simple, but highly efficient distributed system that first uses a small initial segment of data in order to obtain a number of clusters that is proportional to the number of machines, and then has each machine locally and independently apply a local learning algorithm, with no communication between workers at training. In other words, the learning is embarrassingly parallel. See Figure 1 for a schematic representation. Finally, at the prediction time we use a super-fast sublinear algorithm for directing new data points to workers that are relevant for prediction. We show empirically that these techniques can outperform the random partitioning based techniques in terms of accuracy.

For a fixed workload, by increasing the processing power, we obtain a linear speedup. In other words, our paradigm exhibits *strong scaling*.

**Our Contributions** Our main technical contribution is to provide well-founded algorithms for data dependent dispatching, that *cluster* the data, with clusters corresponding to different machines that address challenges arising from distributed learning. For achieving this we consider classic clustering objectives ( $k$ -means,  $k$ -median, and  $k$ -center) and provide algorithms with provable guarantees for partitioning that explicitly incorporate several key important novel constraints that stem from our distributed learning application.

**Balancedness.** We need to make sure that our dispatching procedure balances the data across the different machines. If a machine gets too much data, then it will not be able to use it all; if it gets too little data, then it will not be able to produce an accurate classifier. Thus our clustering algorithm needs to be able to enforce lower and upper bound constraints on the cluster sizes. While prior work has considered upper bounds (this is called capacitated clustering in the literature), much less is known about lower bounds. In fact, as we show in Section 3, lower bounds can result in the optimum solution quality having an unusual behavior in terms of the number of clusters  $k$ , making the problem especially challenging algorithmically.

**Replication.** In order to ensure that our algorithms behave well on points near the boundaries of the clusters, we will need to assign each point to multiple different clusters. This also has the benefit of being more robust to machine failures, highly relevant for distributed settings.

**Efficiency.** We need to be able to perform the clustering based on a small initial random sample and ensure that by finding a clustering that satisfies all our constraints over the sample, we will be able to extend it

over the whole population while maintaining good objective value and also satisfying all the constraints.

In this work we provide the first algorithmic results that simultaneously address all these challenges. Clustering is NP-hard, and adding additional constraints makes this harder, as we shall see in Section 3. For this reason, we devise approximation algorithms with strong theoretical guarantees. Specifically, in Section 2 we provide an algorithm that produces a fault-tolerant clustering that approximately optimizes  $k$ -means,  $k$ -median, and  $k$ -center objectives while also roughly satisfying the given upper and lower bound constraints. At a high level, our algorithm proceeds by first solving a linear program, followed by a careful balance and replication aware rounding scheme. We prove that this algorithm achieves a constant factor approximation to the optimal solution, while violating the replication by a factor of  $2^{\frac{1}{p}}$  and balance constraints to a small  $\frac{p+2}{p}$  factor, where  $p$  is the degree of replication. To our knowledge, no previous work simultaneously seeks to find a clustering with both upper and lower bounds on the cluster sizes, and this can be of independent interest beyond distributed learning.

In Section 3, we provide a number of interesting examples which illustrate the erratic behavior of the clustering objective functions as a function of  $k$ , when there is a lower bound on the optimal cluster sizes. We show there exist clustering instances for which the objective function as a function of  $k$  has an arbitrary number of local maxima. This is in stark contrast to the objective functions with no lower bounds on the cluster sizes, where the objective function with respect to  $k$  is always nonincreasing (even when there are upper bounds on the cluster sizes).

In Section 4, we analyze how many samples we need so that if we only apply our algorithms on a small initial random sample, by using the so-called nearest neighbor extension, we obtain a clustering of the whole population with good objective value that also approximately satisfies all the constraints. Finally, in Section 5, we show the effectiveness of our paradigm over the widely used random partitioning based schemes in several real world image and advertising data sets in terms of performance, and show that our paradigm exhibits Strong Scaling.

**Related Work** The most popular method of dispatch is to do it randomly [34, 35]. This may not produce optimal results because each machine must learn a global model. Another notion is to dispatch the data to pre-determined locations e.g., Yahoo!’s geographically distributed database, PNUTS [12]. However, it does not look at any properties of the data other than physical location.

Previous work in capacitated clustering has focused on upper bounds only [1, 10, 26], while adding lower bounds to the cluster sizes is much less studied [15]. Our algorithm is inspired from the literature, however we solve a much more general and challenging question of simultaneously handling upper and lower bounds on the cluster sizes, and  $p$ -replication.

## 2 Fault Tolerant Balanced Clustering

In this section, we give a single algorithm to handle three clustering objectives with upper and lower bounds on the cluster sizes, and with fault tolerance.

**Setup** A clustering instance consists of a set  $V$  of  $n$  points, and a distance metric  $d$ . Given two points  $i$  and  $j$  in  $V$ , denote the distance between  $i$  and  $j$  by  $d(i, j)$ . The task is to find a set of  $k$  centers  $C = \{c_1, \dots, c_k\}$  and assignments of each point to  $p$  of the centers  $f : V \rightarrow \binom{C}{p}$ , where  $\binom{C}{p}$  represents the subset of  $C^p$  with no duplicates. In this paper, we study three popular clustering objectives:  $k$ -median,  $k$ -means, and  $k$ -center. We focus on the first two, and put the details for  $k$ -center in the Appendix.

- $k$ -median: minimize  $\sum_{i \in V} \sum_{j \in f(i)} d(i, j)$
- $k$ -means: minimize  $\sum_{i \in V} \sum_{j \in f(i)} d(i, j)^2$

We also add size constraints, also known as capacity constraints, so each cluster must have a size between  $n\ell$  and  $nL$ . For simplicity, we assume these values are integral.<sup>2</sup>

<sup>1</sup> If  $p$  is the degree of replication, points may end up with replication between  $p$  and  $p/2$ .

<sup>2</sup> If not, one could replace them by  $\lceil n\ell \rceil$  and  $\lfloor nL \rfloor$  respectively.

1. Find a solution to the following LP relaxation of the clustering IP:

$$\min \sum_{i,j \in V} c_{ij} x_{ij} \quad (\text{LP.1})$$

$$\text{subject to: } \sum_{i \in V} x_{ij} = p, \quad \forall j \in V \quad (\text{LP.2})$$

$$\ell y_i \leq \sum_{j \in V} \frac{x_{ij}}{n} \leq L y_i, \quad \forall i \in V \quad (\text{LP.3})$$

$$\sum_{i \in V} y_i \leq k; \quad (\text{LP.4})$$

$$0 \leq x_{ij} \leq y_i \leq 1, \quad \forall i, j \in V. \quad (\text{LP.5})$$

2. Greedily place points into a set  $\mathcal{M}$  from lowest  $C_j$  to highest (called the “monarchs”), adding a point  $j$  to  $\mathcal{M}$  if it is not within distance  $4C_j$  of any monarch. Partition the point set into coarse clusters (called “empires”) based on the Voronoi partitioning of the monarchs.
3. For each empire  $\mathcal{E}_u$  with total fractional opening  $Y_u \triangleq \sum_{i \in \mathcal{E}_u} y_i$ , give opening  $Y_u / \lfloor Y_u \rfloor$  to the  $\lfloor Y_u \rfloor$  closest points to  $u$ , and give all other points in  $\mathcal{E}_u$  opening 0.
4. Round the  $x_{ij}$ ’s by constructing a minimum cost flow problem on a bipartite graph of centers and points, setting up demands and capacities to handle the bounds on cluster sizes.

**Algorithm 1:** Balanced clustering with fault tolerance

It is well-known that solving the objectives optimally are NP-hard (even without the capacity and fault tolerance generalizations) [18]. We give a bicriteria approximation algorithm for this problem; our algorithm returns a clustering whose cost is at most a constant factor multiple of the optimal solution, while violating the capacity constraints by a small constant factor.

**Algorithm** At a high level, our algorithm proceeds by first solving a linear program, followed by a careful rounding. In particular, we set up an LP whose optimal integral solution is the optimal clustering. We can use an LP solver which will give a fractional solution (for example, the LP may open up  $2k$  ‘half’ centers). Then we greedily pick  $\leq k$  points (called the ‘monarchs’) such that no monarch is within distance  $4C_j$  of another monarch. Then by construction, every non-monarch is within distance  $4C_j$  of a monarch. The empire of a monarch is defined to be its cell in the Voronoi partition over all monarchs. By a Markov inequality, every empire has total opening  $\geq p/2$ , which is at least one for  $p \geq 2$ . Then we merely open the innermost points in the empires as centers, ending with  $\leq k$  centers. Once we have the centers, we find the optimal assignments by setting up a min-cost flow problem. The procedure is summarized in Algorithm 1, and below we provide details, together with the key ideas behind its correctness. The problem is easier for  $p > 1$  because there is more opening in every empire: at least a full center. This allows us to round entirely within each empire and possibly violate the fault-tolerance constraint by a factor of 2 in the rounding.

**Step 1: Linear Program** Now we describe the variables in the integer program that we will relax to an LP. For each  $i$ , let  $y_i$  be an indicator for whether  $i$  is a center. For  $i, j$ , let  $x_{ij}$  be an indicator for whether point  $j$  is assigned to center  $i$ . In the LP, the variables may be fractional, so  $y_i$  represents the fraction to which a center is opened, and  $x_{ij}$  represents the fractional assignment of  $j$  to  $i$ . Let  $(x, y)$  denote an optimal solution to the LP relaxation. For a center  $i$  and point  $j$ , let their contribution to the objective be denoted by  $c_{ij}$ . That is, for  $k$ -median,  $c_{ij} = d(i, j)$  and for  $k$ -means  $c_{ij} = d(i, j)^2$ . Define  $C_j = \sum_i c_{ij} x_{ij}$ , the average cost from point  $j$  to its centers.

It is well-known that the LP in Algorithm 1 has an unbounded integrality gap (the ratio of the optimal LP solution over the optimal integral LP solution), even when the capacities are violated by a factor of  $2 - \epsilon$  [1]. However, with fault tolerance, the integrality is only unbounded when the capacities are violated by a factor of  $\frac{p}{p-1}$ .<sup>3</sup> Intuitively, this is because the  $p$  centers can ‘share’ this violation.

**Step 2: Monarch Procedure** The idea behind step 2 is to partition the points into “empires” such that every point is  $4C_j$  from the center of its empire (the “monarch”), and every empire has total opening at least 1. Then in the next step, we will be able to open at least one center per cluster, and intra-cluster center movements are not too costly. A greedy procedure suffices (for an informal description, see step 2 of Algorithm 1, or for the formal description, see Algorithm 3 in the Appendix). We obtain the following guarantees.

**Lemma 1.** *The output of the monarch procedure satisfies the following properties:*

- (1a) *The clusters partition the point set;*
- (1b) *Each point is close to its monarch:  $\forall j \in \mathcal{E}_u, u \in \mathcal{M}, c_{uj} \leq 4C_j$ ;*
- (1c) *Any two monarchs are far apart:  $\forall u, u' \in \mathcal{M}$  s.t.  $u \neq u', c_{uu'} > 4 \max\{C_u, C_{u'}\}$ ;*
- (1d) *Each empire has a minimum total opening:  $\forall u \in \mathcal{M}, \sum_{j \in \mathcal{E}_u} y_j \geq \frac{p}{2}$ .*

*Proof sketch.* The first three properties follow easily from construction (for property (1c), recall we greedily picked monarchs by the value of  $C_j$ ). For the final property, note that for some  $u \in \mathcal{M}$ , if  $d(i, u) \leq 2C_u$ , then  $i \in \mathcal{E}_u$  (from the triangle inequality and property (1c)). Now, note that  $C_u$  is a weighted average of costs  $c_{iu}$  with weights  $x_{iu}/p$ , i.e.,  $C_u = \sum_i c_{iu} x_{iu}/p$ . By Markov’s inequality, in any weighted average, values greater than twice the average have to get less than half the total weight. That is,

$$\sum_{j: c_{ju} > 2C_u} \frac{x_{ju}}{p} < \sum_{j: c_{ju} > 2C_u} \frac{x_{ju}}{p} \cdot \frac{c_{ju}}{2C_u} < \frac{C_u}{2C_u} = \frac{1}{2}$$

Combining these two facts, for each  $u \in \mathcal{M}$ :

$$\sum_{j \in \mathcal{E}_u} y_j \geq \sum_{j: c_{ju} \leq 2C_u} y_j \geq \sum_{j: c_{ju} \leq 2C_u} x_{ju} \geq \frac{p}{2}. \quad \square$$

**Step 3: Aggregation** The point of this step is to end up with  $\leq k$  centers total. We move the openings so that they are concentrated in the innermost points of each empire. We open these points, and end up with at most  $k$  centers. We show how to accomplish this while violating the capacity constraints by only a factor of  $\frac{p+2}{p}$ . The procedure relies on a suboperation called *Move*, which is the standard way to transfer openings between points (both in  $k$ -median and  $k$ -center arguments [13, 26]) to maintain all LP constraints.

**Definition 1** (Operation “Move”). *The operation “Move” moves a certain opening  $\delta$  from  $a$  to  $b$ . Let  $(x', y')$  be the updated  $(x, y)$  after a movement of  $\delta \leq y_a$  from  $a$  to  $b$ . Define*

$$\begin{aligned} y'_a &= y_a - \delta \\ y'_b &= y_b + \delta \\ \forall u \in V, x'_{au} &= x_{au}(1 - \delta/y_a) \\ \forall u \in V, x'_{bu} &= x_{bu} + x_{au} \cdot \delta/y_a \end{aligned}$$

<sup>3</sup> This is the integrality gap:  $k = 2nL - 1$ , and there are  $nL$  groups of size  $2nL - 1$ . Points in the same group are distance 0, and points in different groups are distance 1. Fractionally, we can open  $2 - \frac{1}{nL}$  facilities in each group to achieve cost 0. But integrally, some group contains at most 1 facility, and thus the capacity violation must be  $2 - \frac{1}{nL}$ . With  $p$  replication, there must be  $p$  centers per group, so the balance violation can be split among the  $p$  centers.

Intuitively, when moving opening from  $y_a$  to  $y_b$ , we update the  $x$ 's so the fractional demand switches from  $a$  to  $b$ . This conserves all LP constraints (with the exception that some variables may become greater than 1) because the constraints are linear (see Section 7 for the details).

The aggregation procedure uses a sequence of *Move*'s to push all fractional openings near each monarch greedily. For an empire  $\mathcal{E}_u$ , let  $Y_u = \sum_{i \in \mathcal{E}_u} y_i$  and  $z_u = \frac{Y_u}{\lfloor Y_u \rfloor}$ . We will give opening  $z_u$  to the  $\lfloor Y_u \rfloor$  closest points to the monarch. Note that by Property (1d), we have  $Y_u \geq 1$  (whenever  $p \geq 2$ ). Then by construction,  $z_u \geq 1$ . In each empire  $\mathcal{E}_u$ , start with the point  $i$  with nonzero  $y_i$  that is farthest away from the monarch  $u$ . Move its opening to the monarch  $u$ . Continue this process until  $u$  has opening exactly  $z_u$ , and then start moving the farthest openings to the point  $j$  closest to the monarch  $u$ . Continue this until the  $\lfloor Y_u \rfloor$  closest points to  $u$  all have opening  $z_u$ . Call the new variables  $(x', y')$ . They have the following properties.

**Lemma 2.** *The aggregated solution  $(x', y')$  satisfies the following constraints:*

(2a) *The opening of each point is either zero or in  $[1, \frac{p+2}{2}]$ :  $\forall i \in V, 1 \leq y'_i < \frac{p+2}{p}$  or  $y'_i = 0$ ;*

(2b) *Each cluster satisfies the capacity constraints:  $i \in V, \ell y'_i \leq \sum_{j \in V} \frac{x'_{ij}}{n} \leq L y'_i$ ;*

(2c) *The total fractional opening is  $k$ :  $\sum_{i \in V} y'_i = k$ ;*

(2d) *Points are only assigned to open centers:  $\forall i, j \in V, x'_{ij} \leq y'_i$ ;*

(2e) *Each point is assigned to  $p$  centers:  $\forall i \in V, \sum_j x'_{ji} = p$ ;*

(2f) *The number of points with non-zero opening is at most  $k$ :  $|\{i \mid y'_i > 0\}| \leq k$ .*

*Proof.* For the first property, recall that each cluster  $\mathcal{E}_u$  has total opening  $\geq \frac{p}{2}$ , so by construction, all  $i$  with nonzero  $y'_i$  has  $y'_i \geq 1$ . We also have  $\frac{Y_u}{\lfloor Y_u \rfloor} \leq \frac{\lfloor Y_u \rfloor + 1}{\lfloor Y_u \rfloor} \leq \frac{p+2}{p}$ , which gives the desired bound.

The next four properties are checking that the LP constraints are still satisfied (except for  $y'_i \leq 1$ ). These follow from the fact that *Move* does not violate the constraints. The last property is a direct result of Properties (2a) and (2c).  $\square$

Now we show that the distance a center moved from the point it serves is bounded by a constant factor.

**Lemma 3.**  *$\forall j \in V$  whose opening moved from  $i'$  to  $i$ ,*

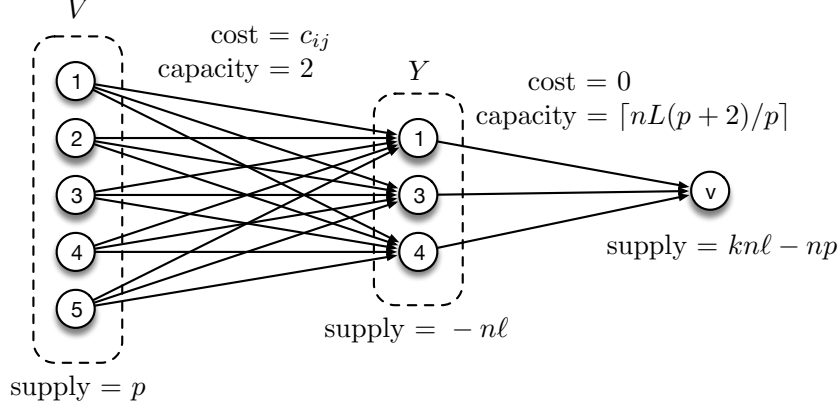
- *$k$ -median:  $d(i, j) \leq 3d(i', j) + 8C_j$ ,*
- *$k$ -means:  $d(i, j)^2 \leq 15d(i', j)^2 + 80C_j$ .*

*Proof.* By construction, if the demand of point  $j$  moved from  $i'$  to  $i$ , then  $\exists u \in \mathcal{M}$  s.t.  $i, i' \in \mathcal{E}_u$  and  $d(u, i) \leq d(u, i')$ . Denote  $j'$  as the closest point in  $\mathcal{M}$  to  $j$ . Then  $d(u, i') \leq d(j', i')$  because  $i' \in \mathcal{E}_u$ . Then,

$$\begin{aligned} d(i, j) &\leq d(i, u) + d(u, i') + d(i', j) \\ &\leq 2d(u, i') + d(i', j) \\ &\leq 2d(j', i') + d(i', j) \\ &\leq 2(d(j', j) + d(j, i')) + d(i', j) \\ &\leq 8C_j + 3d(i', j). \end{aligned}$$

$\square$

We include the proof for  $k$ -means in Appendix 7. Since we have  $\leq k$  points with nonzero opening, we can set them all to 1 to round the  $y$ 's. Now all that is left is to round the  $x$ 's.



**Figure 2: Flow network for rounding the  $x$ 's:** The nodes in each group all have the same supply, which is indicated below each group. The edge costs and capacities are shown above each group. The  $y$ -rounded solution gives a feasible flow in this network. By the Integral Flow Theorem, there exists a minimum cost flow which is integral and we can find it in polynomial time.

**Step 4: Min cost flow** We round the  $x$ 's by setting up a min cost flow problem, where a solution corresponds to an assignment of points to centers. See Figure 2.

We create a bipartite graph with  $V$  on the left (each with supply  $p$ ) and the  $k$  centers on the right (each with demand  $n\ell$ ), and directed edges with weight  $c_{ij}$  and capacity 2. We also add a sink vertex  $v$  with demand  $np - kn\ell$  and directed edges from the centers with weight 0 and capacity  $\frac{p+2}{p}nL$ . This flow problem is carefully set up so that the minimum cost flow that satisfies the capacities corresponds to an optimal clustering assignment. Then using the Integral Flow Theorem, we are guaranteed there is an *integral* assignment that achieves the same optimal cost (and finding the min cost flow is a well-studied polynomial time problem [29]). Thus, we can round the  $x$ 's without incurring any additional cost to the approximation factor.

To put all of these together we have results with all three objectives, which follows directly from Lemma 3 after the  $x$ 's are rounded.

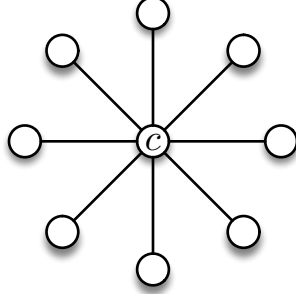
**Theorem 4.** *Algorithm 1 returns an approximate solution for the balanced  $k$ -clustering with  $p$ -replication problem for  $p > 1$ , where the upper capacity constraints are violated by at most a factor of  $\frac{p+2}{p}$ , and each point can be assigned to each center at most twice. The approximation factors are 5, 11, and 95 for  $k$ -center,  $k$ -median, and  $k$ -means, respectively.*

### 3 Structure of Balanced Clustering

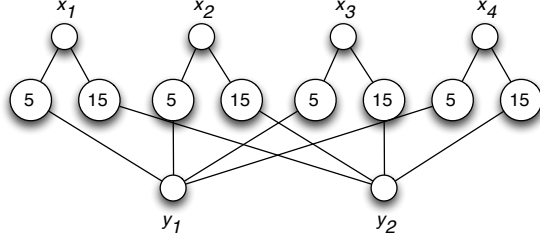
In this section, we explain why adding lower bounds to the cluster sizes is a nontrivial extension. All formal proofs are included in Section 9 of the Appendix, although we provide some proof sketches here.

In uncapacitated clustering, the cost of the optimal clustering is always decreasing in  $k$ . This is easy to see: given an optimal set  $c_1, \dots, c_k$  of centers with cost  $\text{OPT}_k$ , we can make *any* non-center  $p$  into a center, and the resulting clustering has lower cost because at the very least,  $p$  now pays cost 0 and all the other points pay the same cost. There may also be points which are closer to  $p$  than to their center, lowering the objective further. Therefore,  $\text{OPT}_{k+1} \leq \text{cost}(p, c_1, \dots, c_k) \leq \text{OPT}_k$ . This logic extends to the case of capacitated clustering, where the clusters are required to be size  $\leq nL$ .

**Increasing objective function** When we add a lower bound  $n\ell$  on the cluster sizes into the mix, this logic breaks down. Adding a non-center  $p$  may not decrease the cost of the objective, since  $p$ 's closest  $n\ell$



**Figure 3:** A graph in which the objective function strictly increases with  $k$ .



**Figure 4:** Each edge signifies distance 1, and all other distances are 2. The center points are replicated as many times as stated (but each pair of replicated points are still distance 2 away). Finally, add length 1 edges between all pairs in  $x_1, x_2, x_3, x_4, y_1, y_2$ .

points may pay more to connect to  $p$  than to their original center. Consider a star graph with  $n = 10nl + 1$  points (see Figure 3).

The center  $c$  is at distance 1 to the  $10nl$  leaves, and the leaves are at distance 2 from each other. When  $k = 1$ , each point is distance 1 to the center  $c$ . However as we increase  $k$ , the new centers must be leaves, distance 2 from all the other points, so  $nl - 1$  points must pay 2 instead of 1 for each extra center (see Lemma 22 in the Appendix). It is also easy to achieve an objective that strictly decreases up to a local minimum  $k'$ , and then strictly increases onward, by adding  $k'$  copies of the center of the star.

Note for this example, the problem goes away if we are allowed to place multiple centers on a single point (in the literature, this is called “soft capacities”, as opposed to enforcing one center per point, called “hard capacities”). We can create a stronger example as follows. We make  $m$  groups of clusters, each size  $2nl - 1$ . Two points are distance 0 if they are in the same group, otherwise 1. Then when  $k = m$ , we can put one center per group and achieve cost 0, but for  $k = m + 1$ , there must be two centers in one group, so the cost is 1 (see Lemma 23 in the Appendix).

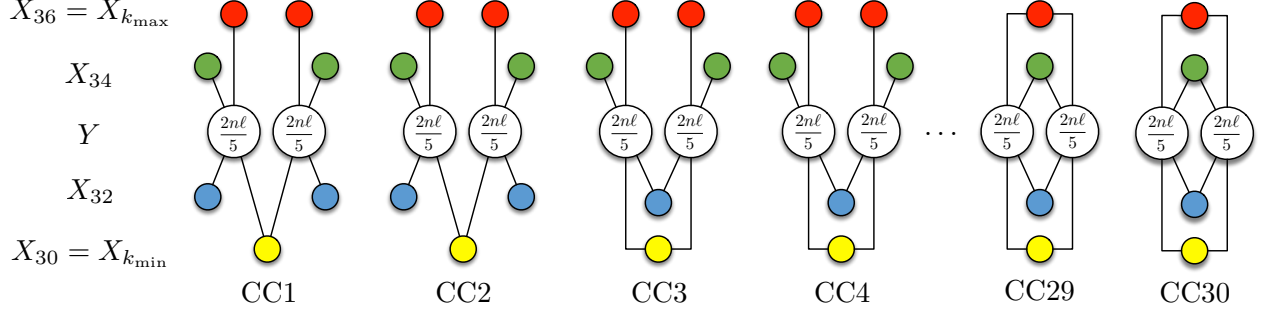
**Local maxima** So far, we have seen examples in which the objective decreases with  $k$ , until it hits a minimum (where capacities start to become violated), and then the objective strictly increases. The next natural question to ask, is whether the objective can also have a local maximum. It turns out, this is possible, but we must start using more complicated constructions.

**Lemma 5.** *There exists a clustering instance in which the objective function contains a local maximum with respect to  $k$ , for  $k$ -center,  $k$ -median, and  $k$ -means.*

*Proof sketch.* Consider the graph in Figure 4, and let  $nl = 21$ . Note that, as every distance is either 1 or 2, the triangle inequality is trivially satisfied, so the point set is a metric. It is easy to verify that  $k = 2$  and  $k = 4$  have valid clusterings using only length 1 edges, using centers  $\{y_1, y_2\}$  and  $\{x_1, x_2, x_3, x_4\}$ , respectively.

Now consider  $k = 3$ . The crucial property is that by construction,  $y_1$  and any  $x_i$  cannot simultaneously be centers and each satisfy the capacity to distance 1 points, because the union of their distance 1 neighborhoods





**Figure 5:** An example when  $m = 3$ . Each  $X_k$  is a different color. The white circles represent  $2n\ell/5$  points from  $Y$  each.

is less than  $2n\ell$ . After this observation, it is straightforward to check that no set of 3 centers can achieve a clustering with only distance 1 edges.  $\square$

The previous example does not work for the case of soft capacities, since the set of centers  $\{x_1, y_2, y_2\}$  allows every point to have an edge to its center. Now we come to our main theorem of this section. With a careful construction, we are able to generalize the previous theorem to achieve *any number* of local maxima, even for soft capacities.

**Theorem 6.** *For all  $m \in \mathbb{N}$ , there exists a clustering instance with  $m$  local maxima, for  $k$ -center,  $k$ -median, and  $k$ -means, even for soft capacities.*

*Proof sketch.* As in the previous lemma, we will construct a set of points in which each pair of points are either distance 1 or 2. It is convenient to define a graph on the set of points, in which an edge signifies a distance of 1, and all non-edges denote distance 2. We will construct a clustering instance where the objective value for all even values of  $k$  between  $10m$  and  $12m$  is low and the objective value for all odd values of  $k$  between  $10m$  and  $12m$  is high. The  $m$  odd values will be the local maxima. We will set the lower bound  $n\ell$  to be the product of all the even integers between  $10m$  and  $12m$ .

We start by creating a distinct set of “good” centers,  $X_k$ , for each even value of  $k$  between  $10m$  and  $12m$ . Let  $X$  be the union of these sets. The set  $X_k$  contains  $k$  points which will be the optimal centers for a  $k$ -clustering in our instance. Then we will add an additional set of points,  $Y$ , and add edges from  $Y$  to the centers in  $X$  with the following properties.

1. For each even value of  $k$  between  $10m$  and  $12m$ , there is an assignment of the points in  $Y$  to the centers in  $X_k$  so that points in  $Y$  are only assigned to adjacent centers and the capacity constraints are satisfied.
2. Each of the good centers in  $X$  is adjacent to no more than  $\frac{6}{5} \cdot n\ell$  points in  $Y$ .
3. For each good center  $x$  in  $X_k$ , there is at least one point  $x'$  in every other set  $X_{k'}$  (for  $k' \neq k$ ) so that the number of points in  $Y$  that are adjacent to both  $x$  and  $x'$  is at least  $\frac{2}{5} \cdot n\ell$ .
4. Any subset of the centers in  $X$  that does not contain any complete set of good centers  $X_k$  for some even  $k$  is non-adjacent to at least one point in  $Y$ .

Whenever we add a point to  $Y$ , we give it an edge to exactly one point from each  $X_k$ . This ensures that each  $X_k$  partitions  $Y$ . We first create connected components as in Figure 5 that each share  $\frac{2}{5} \cdot n\ell$  points from  $Y$ , to satisfy Property 3.

For property 4, we add one additional point to  $Y$  for every combination of picking one point from each  $X_k$ . This ensures that any set which does not contain at least one point from each  $X_k$  will not be a valid

partition for  $Y$ . Note that in the previous two steps, we did not give a single center more than  $\frac{6}{5} \cdot n\ell$  edges, satisfying property 2. Then we add “filler” points to bring every center’s capacity up to at least  $n\ell$ , which satisfies property 1.

Now we explain why properties 1-4 are sufficient to finish off the proof. Property 1 guarantees that the for each even value of  $k$  there is a clustering where the cost of each point in  $Y$  is one, which results in a good clustering objective.

Properties 2 and 3 guarantee that any set including a full  $X_k$  and a point from a different  $X_{k'}$  cannot achieve cost 1 for each point without violating the capacities. Property 4 guarantees that any set without a full  $X_k$  cannot achieve cost 1 for each point. This completes the proof.  $\square$

## 4 Properties of Nearest Neighbor Extension

The next step is to extend a clustering on the the sample  $S = \{x_1, \dots, x_n\}$  to the distribution  $\mu$  to get good objective value and satisfy cluster size constraints. Key challenges are to ensure that balance constraints remain approximately satisfied, and to bound the loss in quality due to the fact that our algorithm only chooses clusterings from a restricted class of clusterings that depends on the set  $S$ .

**Informal Description** Without capacity constraints, we can easily extend a clustering of the set  $S$  to the entire space  $\mathcal{X}$  by assigning each point  $x$  to its nearest  $p$  centers. For any given set of centers, this is the optimal cluster assignment [8]. But this is no longer true in the realm of capacitated clustering, since to satisfy the capacity constraints, points may not be assigned to their nearest  $p$  centers. Independently of how we cluster the set  $S$ , we extend the clustering to the entire space  $\mathcal{X}$  using the nearest neighbor extension, similar in spirit to [9]. Each data point  $x_i \in S$  acts as a representative for the set of points closer to it than any other data point, denoted by  $V_i = \{x \in \mathcal{X} : \text{NN}_S(x) = x_i\}$ , where  $\text{NN}_S(x) = \text{argmin}_{x' \in S} d(x, x')$  is the nearest neighbor in the set  $S$  to the point  $x$ . The sets  $V_i$  for  $i = 1, \dots, n$  are the cells in the Voronoi partition of  $\mathcal{X}$  induced by the sample  $S$ . We extend a clustering of  $S$  to the entire space  $\mathcal{X}$  by assigning each point in the Voronoi tile  $V_i$  to the same cluster centers as the sample point  $x_i$  for  $i = 1, \dots, n$ .

When we assign the point  $x_i$  to some cluster, we implicitly also assign all the points in  $V_i$  to that cluster. If a Voronoi tile  $V_i$  has high probability mass, then the point  $x_i$  is more influential on the clustering objective and balance constraints for the nearest neighbor extension. Therefore, when we cluster the set  $S$ , we consider a weighted version of capacitated clustering where each point  $x_i$  is weighted by the mass of  $V_i$ . Since the probability distribution  $\mu$  is unknown, we estimate the weight of each  $V_i$  by counting the number of points  $n_i$  from a second sample  $S'$  drawn randomly from  $\mu$  that land in  $V_i$ . Pseudocode for our procedure is given in Algorithm 2.

**Notation** Let  $\mathcal{X}$  be a set of bounded diameter  $D$ . Each clustering of  $\mathcal{X}$  can be represented as a pair  $(f, c)$  where  $f : \mathcal{X} \rightarrow \binom{[k]}{p}$  represents the cluster assignments and  $c \in \mathcal{X}^k$  is a list of centers. The population-level  $k$ -median objective is the expected total distance from a point  $x$  sampled from  $\mu$  to its  $p$  assigned centers:

$$Q(f, c) = \mathbb{E}_{x \sim \mu} \left[ \sum_{i \in f(x)} d(x, c(i)) \right].$$

The population-level capacity constraints require that the probability mass of each cluster  $i$ ,  $\mathbb{P}_{x \sim \mu}(i \in f(x))$ , is between  $\ell$  and  $L$ . Similarly, a clustering of the data set  $S$  is a pair  $(g, c)$  for some assignment  $g : S \rightarrow \binom{[k]}{p}$  and centers represented by  $c$ . The weight for point  $x_i$  is  $w_i = \mathbb{P}_{x \sim \mu}(\text{NN}_S(x) = x_i)$ . The weighted  $k$ -median objective on  $S$  is

$$Q_n(g, c) = \sum_{j=1}^n w_j \sum_{i \in g(x_j)} d(x_j, c(i))$$

**Parameters:**  $k, p, \ell, L$ , second sample size  $n'$ .

**procedure** NNEXTENSION( $S = \{x_1, \dots, x_n\}$ )

    Draw  $S'$  of size  $n'$  iid from  $\mu$ .

    Estimate  $\hat{w}_i = |S' \cap V_i|/n'$ .

    Compute  $(g_n, c_n)$  by minimizing  $\hat{Q}_n(g, c)$  subject to size constraints  $(\ell, L)$ . **return**  $(\bar{g}_n, c_n)$ .

**end procedure**

**Algorithm 2:** Nearest Neighbor Clustering Extension

The weighted capacity constraints require that the total weight of each cluster  $i$ ,  $\sum_{j:i \in g(x_j)} w_j$  is between  $\ell$  and  $L$ . Given estimates  $\hat{w}_1, \dots, \hat{w}_n$  of the true weights, the estimated objective function is

$$\hat{Q}_n(g, c) = \sum_{j=1}^n \hat{w}_j \sum_{i \in g(x_j)} d(x_j, c(i)),$$

and the estimated weight of a cluster is  $\sum_{j:i \in g(x_j)} \hat{w}_j$ . Finally, for any clustering  $(g, c)$  of  $S$ , define the nearest neighbor extension to be  $(\bar{g}, c)$  where  $\bar{g}(x) = g(\text{NN}_S(x))$ .

Before stating our main result, we first show that if we take the second sample size  $n'$  to be  $\tilde{O}(n/\epsilon^2)$ , then with high probability the error in any sum of the estimated weights  $\hat{w}_j$  is at most  $\epsilon$ .

**Lemma 7.** *For any  $\epsilon > 0$  and  $\delta > 0$ , if we set  $n' = O(\frac{1}{\epsilon^2}(n + \log \frac{1}{\delta}))$  in Algorithm 2, then with probability at least  $1 - \delta$  we have  $|\sum_{i \in I} (w_i - \hat{w}_i)| \leq \epsilon$  uniformly for all index sets  $I \subset [n]$ .*

*Proof.* For any index set  $I \subset [n]$ , let  $V_I$  denote the union  $\bigcup_{i \in I} V_i$ . Since the sets  $V_1, \dots, V_n$  are disjoint, for any index set  $I$  we have that  $\mu(V_I) = \sum_{i \in I} w_i$  and  $\hat{\mu}(V_I) = \sum_{i \in I} \hat{w}_i$ , where  $\hat{\mu}$  is the empirical measure induced by the second sample  $S'$ . Therefore it suffices to show uniform convergence of  $\hat{\mu}(V_I)$  to  $\mu(V_I)$  for the  $2^n$  index sets  $I$ . Applying Hoeffding's inequality to each index set and the union bound over all  $2^n$  index sets, we have that

$$\mathbb{P}\left(\sup_{I \subset [n]} \left| \sum_{i \in I} w_i - \hat{w}_i \right| > \epsilon\right) \leq 2^n e^{-2n'\epsilon^2}.$$

Setting  $n' = O(\frac{1}{\epsilon^2}(n + \log \frac{1}{\delta}))$  results in the right hand side being equal to  $\delta$ .  $\square$

Next we relate the weighted capacity constraints and objective over the set  $S$  to the population-level constraints and objective.

**Lemma 8.** *Let  $(g, c)$  be any clustering of  $S$  that satisfies the weighted capacity constraints with parameters  $\ell$  and  $L$ . Then the nearest neighbor extension  $(\bar{g}, c)$  satisfies the population-level capacity constraints with the same parameters and the difference between the sample and population objectives is bounded as follows:*

$$|Q_n(g, c) - Q(\bar{g}, c)| \leq p \mathbb{E}_{x \sim \mu} [d(x, \text{NN}_S(x))].$$

*Proof.* The fact that  $\bar{g}$  satisfies the population-level capacity constraints follows immediately from the definition of the weights  $w_1, \dots, w_n$ .

By the triangle inequality, the population-level objective can be bounded as

$$Q(\bar{g}, c) \leq \mathbb{E}_{x \sim \mu} \left[ \sum_{i \in \bar{g}(x)} d(x, \text{NN}_S(x)) \right] + \mathbb{E}_{x \sim \mu} \left[ \sum_{i \in \bar{g}(x)} d(\text{NN}_S(x), c(i)) \right] = p \mathbb{E}_{x \sim \mu} [d(x, \text{NN}_S(x))] + Q_n(g, c).$$

Similarly, we have that  $Q_n(g, c) \leq p \mathbb{E}_{x \sim \mu} [d(x, \text{NN}_S(x))] + Q(\bar{g}, c)$ , finishing the proof.  $\square$

**Main Result** We bound the sub-optimality of the clustering  $(\bar{g}_n, c_n)$  returned by Algorithm 2 with respect to any clustering  $(f^*, c^*)$  of the population. The bound will depend on

1. the quality of the finite-data algorithm,
2. the “average radius” of the Voronoi cells  $\alpha(S) = \mathbb{E}_{x \sim \mu}[d(x, \text{NN}_S(x))]$ , and
3. the bias from returning clusterings that are constant over  $V_i$ ,

$$\beta(S, \ell, L) = \min_{h, c} \{Q(\bar{h}, c) - Q(f^*, c^*) \mid h \text{ satisfies balance constraints } (\ell, L)\},$$

where the minimum is taken over all clusterings  $(h, c)$  of the sample  $S$  and  $(\bar{h}, c)$  denotes the nearest neighbor extension. When  $\ell, L$  are clear from the context, we just write it as  $\beta(S)$ .

**Theorem 9.** For any  $\epsilon > 0, \delta > 0$ , let  $(\bar{g}_n, c_n)$  be the output of Algorithm 2 with parameters  $k, p, \ell, L$  and second sample size  $n' = O((n + \log 1/\delta)/\epsilon^2)$ . Let  $(f^*, c^*)$  be any clustering of  $\mathcal{X}$  and  $(g_n^*, c_n^*)$  be an optimal clustering of  $S$  under  $\hat{Q}_n$  satisfying the estimated weighted balance constraints  $(\ell, L)$ . Suppose the finite data algorithm used satisfies  $\hat{Q}(g_n, c_n) \leq r \cdot \hat{Q}(g_n^*, c_n^*) + s$ . Then w.p.  $\geq 1 - \delta$  over the second sample the output  $(\bar{g}_n, c_n)$  will satisfy the balance constraints with  $\ell' = \ell - \epsilon$  and  $L' = L + \epsilon$  and we have

$$Q(\bar{g}_n, c_n) \leq r \cdot Q(f^*, c^*) + s + 2(r+1)pD\epsilon + p(r+1)\alpha(S) + r\beta(S, \ell + \epsilon, L - \epsilon).$$

*Proof.* Lemma 7 guarantees that when the second sample is of size  $O(\frac{1}{\epsilon^2}(n + \log \frac{1}{\delta}))$  then with probability at least  $1 - \delta$ , for any index set  $I \subset [n]$ , we have  $|\sum_{i \in I} w_i - \hat{w}_i| \leq \epsilon$ . For the remainder of the proof, assume that this high probability event holds.

First we argue that the clustering  $(g_n, c_n)$  satisfies the true weighted capacity constraints with the slightly looser parameters  $\ell' = \ell - \epsilon$  and  $L' = L + \epsilon$ . Since the clustering  $(g_n, c_n)$  satisfies the estimated weighted capacity constraints, the high probability event guarantees that it will also satisfy the true weighted capacity constraints with the looser parameters  $\ell' = \ell - \epsilon$  and  $L' = L + \epsilon$ . Lemma 8 then guarantees that the extension  $(\bar{g}_n, c_n)$  satisfies the population-level capacity constraints with parameters  $\ell'$  and  $L'$ .

Next we bound the difference between the estimated and true weighted objectives for any clustering  $(g, c)$  of  $S$ . For each point  $x_j$  in the set  $S$ , let  $C_j = \sum_{i \in g(x_j)} d(x_j, c(i))$  be the total distance from point  $x_j$  to its  $p$  assigned centers under clustering  $(g, c)$ , and let  $J$  be the set of indices  $j$  for which  $\hat{w}_j > w_j$ . Then by the triangle inequality

$$\begin{aligned} |\hat{Q}_n(g, c) - Q_n(g, c)| &\leq \left| \sum_{j \in J} (\hat{w}_j - w_j) C_j \right| + \left| \sum_{j \notin J} (w_j - \hat{w}_j) C_j \right| \\ &\leq \left( \left| \sum_{j \in J} (\hat{w}_j - w_j) \right| + \left| \sum_{j \notin J} (w_j - \hat{w}_j) \right| \right) pD \\ &\leq 2pD\epsilon, \end{aligned} \tag{2}$$

where the second inequality follows from the fact that  $C_j \leq pD$  and the sum has been split so that  $(\hat{w}_j - w_j)$  is always positive in the first sum and negative in the second.

Finally, let  $(h_n, c'_n)$  be the clustering of  $S$  that attains the minimum in the definition of  $\beta(S, \ell + \epsilon, L - \epsilon)$ . That is, the clustering of  $S$  satisfying the capacity constraints with parameters  $\ell + \epsilon$  and  $L - \epsilon$  whose nearest neighbor extension has the best population-level objective. Then combining Lemma 8, equation (2), the approximation guarantees for  $(g_n, c_n)$  with respect to  $\hat{Q}_n$ , and the optimality of  $(g_n^*, c_n^*)$ , we have the

following:

$$\begin{aligned}
Q(\bar{g}_n, c_n) &\leq Q_n(g_n, c_n) + p\alpha(S) \\
&\leq \hat{Q}_n(g_n, c_n) + 2pD\epsilon + p\alpha(S) \\
&= \hat{Q}_n(g_n, c_n) - r \cdot \hat{Q}_n(g_n^*, c_n^*) + r \cdot \hat{Q}_n(g_n^*, c_n^*) + 2pD\epsilon + p\alpha(S) \\
&\leq s + 2pD\epsilon + p\alpha(S) + r \cdot \hat{Q}_n(h_n, c'_n) \\
&\leq s + 2(r+1)pD\epsilon + p\alpha(S) + r \cdot Q_n(h_n, c'_n) \\
&\leq s + 2(r+1)pD\epsilon + p(r+1)\alpha(S) + r \cdot Q(\bar{h}_n, c'_n) \\
&\leq s + 2(r+1)pD\epsilon + p(r+1)\alpha(S) + r \cdot \beta(S, \ell + \epsilon, L - \epsilon) + r \cdot Q(f^*, c^*),
\end{aligned}$$

concluding the proof.  $\square$

The above theorem applies for any set  $S$ , but the quality of the bound depends on  $\alpha(S)$  and  $\beta(S)$ , which measure how well the set  $S$  represents the distribution  $\mu$ . We now bound  $\alpha(S)$  and  $\beta(S)$  when  $S$  is a large enough iid sample drawn from  $\mu$  under various conditions on  $\mu$  and the optimal clustering. The proofs can be found in Section 10 of the appendix.

**Bounding  $\alpha(S)$**  We bound the sample size required to make  $\alpha(S)$  small when  $\mathcal{X} \subseteq \mathbb{R}^q$  and  $S$  is drawn randomly from an arbitrary  $\mu$ . Additionally, when the distribution has a lower intrinsic dimension, we can do better. The doubling condition is one such a condition. Let  $B(x, r)$  be a ball of radius  $r$  around  $x$  with respect to the metric  $d$ . A measure  $\mu$  with support  $\mathcal{X}$  is said to be a doubling measure of dimension  $d_0$  if for all points  $x \in \mathcal{X}$  and all radii  $r > 0$  we have  $\mu(B(x, 2r)) \leq 2^{d_0} \mu(B(x, r))$ .

**Lemma 10.** *For any  $\epsilon, \delta > 0$ , and  $\mathcal{X} \subseteq \mathbb{R}^q$ , if a randomly drawn  $S$  from  $\mu$  is of size  $O(q^{q/2} \epsilon^{-(q+1)} (q \log \frac{\sqrt{q}}{\epsilon} + \log \frac{1}{\delta}))$  in the general case, or  $O(\epsilon^{-d_0} (d_0 \log \frac{1}{\epsilon} + \log \frac{1}{\delta}))$  if  $\mu$  is doubling with dimension  $d_0$ , then w.p  $\geq 1 - \delta$ ,  $\alpha(S) \leq \epsilon D$*

**Bounding  $\beta(S)$**  Bubeck et al. [9] provide a worst case lower bound when  $f^*$  is continuous almost everywhere. Again, one can do better for well-behaved input. The Probabilistic Lipschitzness (PL) condition [31, 32] says that  $f$  is  $\phi$ -PL if the probability mass of points that have non-zero mass of differently labeled points in a  $\lambda D$ -ball around them is at most  $\phi(\lambda)$ . If a clustering function  $f$  is PL, it means the clusters are, in some sense, “round”- that the probability mass “close to” the boundaries of the clusters is small. Under this condition, we have the following sample complexity result for  $\beta$ . We can compare to a clustering with slightly tighter size constraints:

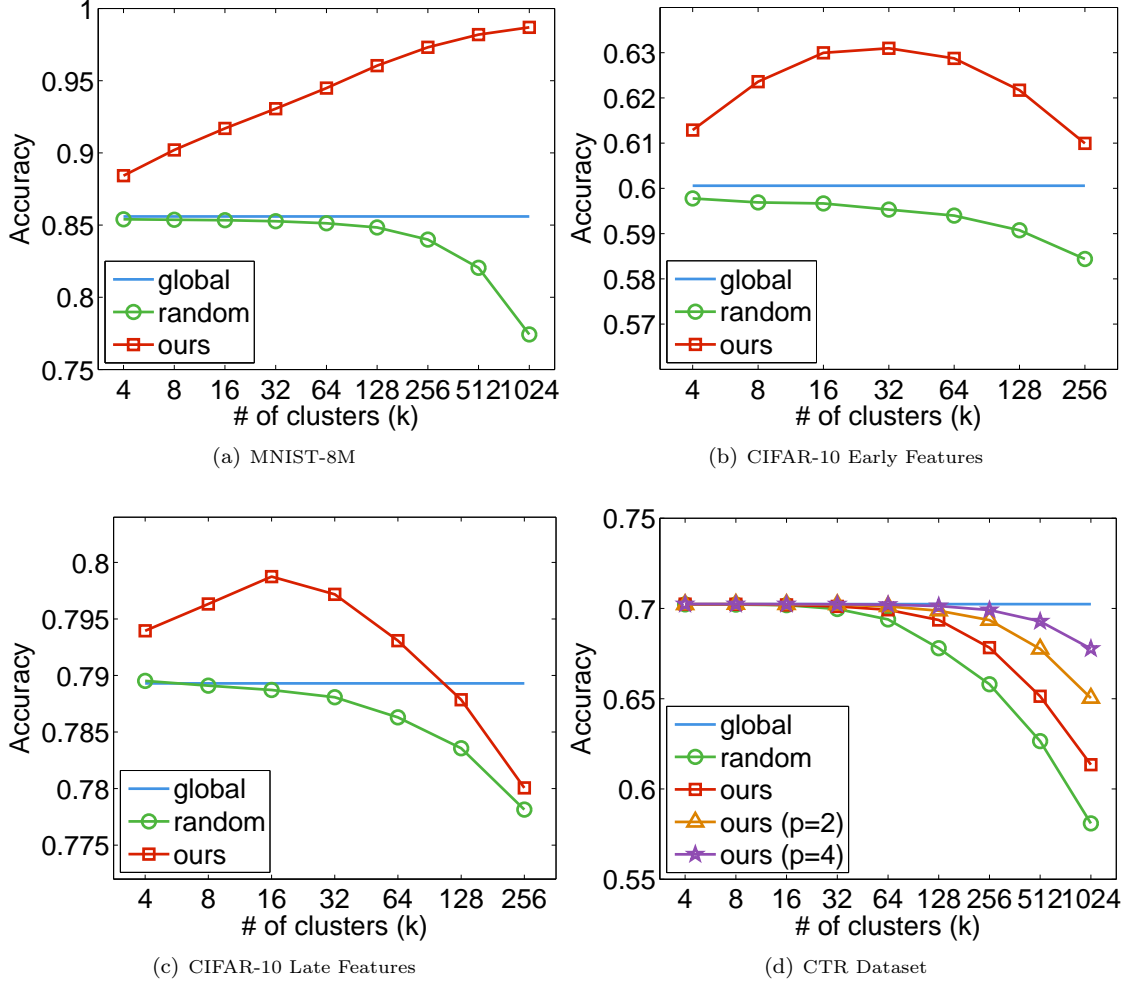
**Lemma 11.** *Let  $\mu$  be a measure on  $\mathbb{R}^q$  with support  $\mathcal{X}$  of diameter  $D$ . Let  $f^*$ , some clustering of  $\mu$  that satisfies capacities  $(\ell + \epsilon, L - \epsilon)$ , be  $\phi$ -PL. If we see a sample  $S$  drawn iid from  $\mu$  of size  $O\left(\frac{1}{\epsilon} \left(\frac{1}{\phi^{-1}(\epsilon/2)}\right)^q (q \log \frac{\sqrt{q}}{\phi^{-1}(\epsilon/2)} + \log \frac{1}{\delta})\right)$  in the general case or  $O\left(\left(\frac{1}{\phi^{-1}(\epsilon)}\right)^{d_0} (d_0 \log \frac{4}{\phi^{-1}(\epsilon)} + \log \frac{1}{\delta})\right)$  when  $\mu$  is a doubling measure of dimension  $d_0$  then, w.p. at least  $1 - \delta$  over the draw of  $S$ , we have that  $\beta(S, \ell, L) \leq pD\epsilon$ .*

## 5 Experiments

In this section we compare the performance of our distributed learning paradigm against two common alternative schemes:

**Global** Workers update on the shared model with synchronization. Equivalently, the model is obtained by having all the data on a single machine.

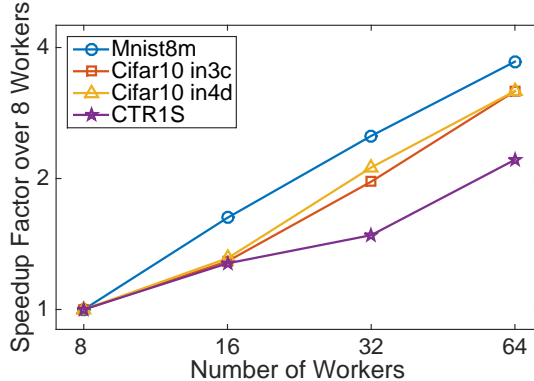
**Random** Data is randomly partitioned to workers, where each worker learns independently from a subset of the data.



**Figure 6:** Study of classification accuracy vs number of clusters,  $k$  for MNIST-8M, CIFAR-10 (with two different feature representations), and a CTR dataset. The s.d. over different runs is  $\sim 10^{-3}$  and therefore omitted.

In particular, we study the performance of our method and the two methods above for various values of  $k$ , the number of clusters. Moreover, for a fixed  $k$ , we analyze the scalability of our method by observing the speedup obtained by increasing the processing power, i.e. number of machines used.

**Implementation** We used an approximate implementation of our algorithm. First, an initial sample is drawn uniformly at random. We cluster this sample using  $k$ -means++ [3]. We use the popular  $k$ -means++ algorithm in order to scale to these large datasets considered here. We expect that using our clustering algorithms instead would further improve performance. For fault tolerance, we use a modified Lloyd’s iteration that maintains  $p$  assignments for each data point. To satisfy the capacity constraints, we use two simple heuristics: while the smallest cluster violates the lower capacity constraint, the smallest cluster is merged with the cluster whose center is closest to its own, and the center is updated as in Lloyd’s algorithm. After this step, there are possibly fewer than  $k$  clusters, but they all satisfy the lower capacity. Any cluster that violates the upper capacity constraints is randomly partitioned into evenly-sized clusters that satisfy the upper capacity constraint. This clustering approach may produce fewer or more than  $k$  clusters. In all



**Figure 7:** Linear speedup: When the number of workers is doubled, the time for dispatch, learning and testing (averaged over 5 runs) drops by a constant factor.  $k$  was set to 128 for CIFAR-10 (both), CTR and 512 for MNIST-8M

experiments, we use the capacity constraints  $\ell = p/(2k)$  and  $L = 2p/k$ . Finally, rather than implementing exact nearest neighbor dispatch as in Algorithm 2, we use the random partition tree (RPT) algorithm of Dasgupta and Sinha [14] to dispatch based on approximately nearest neighbors to achieve huge speed-ups. Further, we approximate  $w_i$ 's, the weights of Voronoi partitions, by their expected values,  $1/n$ . The julia code is available online.

For the classification, we used the linear one-vs-all multi-class SVM provided by Liblinear [16]. The regularization parameter is chosen by 5-fold cross validation.

During deployment, given a new test point, in a real system, one would dispatch the point randomly to one of its  $p$  clusters and use the prediction made by that cluster. In our experiments, we measure the *expected accuracy* over the choice the cluster. The RPT based dispatcher sends the point to all appropriate clusters. Predictions are made independently in each cluster using the learnt models. This test point is assigned an accuracy score of number of correct predictions divided by the total number of predictions. The accuracy on the entire dataset is calculated as an average of these scores.

**Experimental Setup** We now describe the distributed implementation used for the experiments. We start one worker process on each of the available processing cores. First, a single worker subsamples the data, clusters the subsample into  $k$  clusters, and then builds a random partition tree for fast nearest neighbor lookup. The subsample, clustering, and random partition tree describe a dispatching rule, which is then copied to every worker. Training the system has two steps: first, the training data is dispatched to the appropriate workers, and then each worker learns a model for the clusters they are responsible for. During the deployment phase, the workers load the training data in parallel and send each example to the appropriate workers (as dictated by the dispatch rule). To minimize network overhead examples are only sent over the network in batches of 5000. During the training phase, each worker calls Liblinear to learn a model for each cluster they were responsible for. For testing, the testing data is loaded in parallel by the workers and the appropriate workers are queried for predictions. The experiments were performed on cluster of 15 machines, each with 8 Intel(R) Xeon(R) cores of clock rate 2.40 GHz and 32GB shared memory per machine.

**Datasets** We used three public datasets ranging from images to text to evaluate the these methods.

**MNIST-8M:** We used the raw pixels of this handwritten image dataset [27], which has 8 million examples and 784 features.

**CIFAR-10:** The CIFAR-10 dataset [23] is an image classification task with 10 classes. Following Krizhevsky *et al.* [24] we include 50 copies of each training, example each randomly rotated and cropped to get

a training set of 2.5 million examples. We extracted the features from the Google Inception<sup>4</sup> [30] by using the output of an early layer (layer in3c) and a later layer (layer in4d).

**CTR:** The CTR dataset contains ad impressions from a commercial search engine and the label is whether or not it was clicked by a user. It has 860K examples with 232 continuous-valued features.

**Results** The results are shown in Figure 6. As can be seen, our method always performs better than random partitioning. The performance gap varies over datasets. For the simplest MNIST-8M, our method achieves tremendous improvements in performance (about 14%) compared to the random partitioning scheme. But the improvements are not as pronounced when the clustering structure is less obvious.

Also note that there is an optimal number of clusters,  $k^*$  for each dataset, where the classification accuracy obtained is the highest. It is large for simple datasets ( $k^* \geq 1024$  for MNIST-8M while  $k^* = 16$  for CTR). Further, there seems to be a “safe” number of clusters where the performance of our paradigm is at least as good as the global linear model. This value is larger than 16 for all datasets. That is, the proposed method can enjoy the embarrassingly parallel training without any loss of accuracy.

It is also interesting to note that the performance improves with increasing  $p$ . This suggests that we ought to replicate points not just for fault-tolerance, but also for better performance.

Our system exhibits the desirable property of *Strong Scaling*, as demonstrated by Figure 7. That is, for a fixed workload, if we have twice as much processing power, the time for dispatch, training and deployment roughly drops by a constant factor, roughly until the number of worker processes equals the number of clusters,  $k$ . We get no further benefit after this point because each cluster is only served by a single worker. In principle, one could have multiple workers catering to a single cluster.

Overall, the results support our claim that data dependent partitioning scheme is good in both theory and practice.

## 6 Conclusion

To sum up, we propose a novel data-dependent resource allocation paradigm that could also be of interest beyond distributed learning. We cast this as a clustering problem with additional constraints for load balancing and fault tolerance. Balanced clustering is challenging, and the optimal solution can behave weirdly, as described in 3. We were able to overcome these challenges, showing a polynomial time LP-rounding algorithm with strong approximation guarantees, and this is of independent interest beyond the distributed learning. Moreover, we propose nearest neighbor extensions for dispatch and give sample complexity results for good performance. On the experimental side, a large-scale implementation on three different datasets shows that data dependent partitioning is also very effective in practice. After performing data dependent partitioning schemes as described, one can go beyond communication free learning and explore how allowing some amount of communication can help improve performance.

## References

- [1] Karen Aardal, Pieter L van den Berg, Dion Gijswijt, and Shanfei Li. Approximation algorithms for hard capacitated k-facility location problems. *European Journal of Operational Research*, (2):358–368, 2015.
- [2] Hyung-Chan An, Aditya Bhaskara, Chandra Chekuri, Shalmoli Gupta, Vivek Madan, and Ola Svensson. Centrality of trees for capacitated k-center. In *Integer Programming and Combinatorial Optimization*, pages 52–63. Springer, 2014.

---

<sup>4</sup>For the specific network structure, refer to <https://github.com/dmlc/mxnet/blob/master/example/notebooks/cifar-recipe.ipynb>.



- [3] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.
- [4] Maria-Florina Balcan, Avrim Blum, Shai Fine, and Yishay Mansour. Distributed learning, communication complexity, and privacy. In *Conference on Learning Theory*, 2012.
- [5] Maria-Florina Balcan, Steven Ehrlich, and Yingyu Liang. Distributed k-means and k-median clustering on general communication topologies. In *Advances in Neural Information Processing Systems*, 2013.
- [6] Maria-Florina Balcan, Vandana Kanchanapally, Yingyu Liang, and David Woodruff. Improved distributed principal component analysis. In *Advances in Neural Information Processing Systems*, 2014.
- [7] J. Barilan, G. Kortsarz, and D. Peleg. How to allocate network centers. *Journal of Algorithms*, 15(3):385–415, 1993.
- [8] Shai Ben-David. A framework for statistical clustering with constant time approximation algorithms for k-median and k-means clustering. *Machine Learning*, 66(2):243–257, 2007.
- [9] Sébastien Bubeck and Ulrike von Luxburg. Nearest neighbor clustering: A baseline method for consistent clustering with arbitrary objective functions. *The Journal of Machine Learning Research*, 10:657–698, 2009.
- [10] Jarosław Byrka, Krzysztof Fleszar, Bartosz Rybicki, and Joachim Spoerhase. Bi-factor approximation algorithms for hard capacitated k-median problems. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 722–736. SIAM, 2015.
- [11] Moses Charikar, Sudipto Guha, Éva Tardos, and David B Shmoys. A constant-factor approximation algorithm for the k-median problem. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 1–10. ACM, 1999.
- [12] Brian F Cooper, Raghu Ramakrishnan, Utkarsh Srivastava, Adam Silberstein, Philip Bohannon, Hans-Arno Jacobsen, Nick Puz, Daniel Weaver, and Ramana Yerneni. Pnuts: Yahoo!’s hosted data serving platform. *Proceedings of the VLDB Endowment*, 1(2):1277–1288, 2008.
- [13] Marek Cygan, MohammadTaghi Hajiaghayi, and Samir Khuller. Lp rounding for k-centers with non-uniform hard capacities. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 273–282. IEEE, 2012.
- [14] Sanjoy Dasgupta and Kaushik Sinha. Randomized partition trees for exact nearest neighbor search. *Algorithmica*, 72(1):237–263, 2015.
- [15] Alina Ene, Sarel Har-Peled, and Benjamin Raichel. Fast clustering with lower bounds: No customer too far, no shop too small. *CoRR*, abs/1304.7318, 2013.
- [16] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.
- [17] Sudipto Guha, Adam Meyerson, and Kamesh Munagala. Hierarchical placement and network design problems. In *FOCS*, pages 603–612. IEEE Computer Society, 2000.
- [18] Kamal Jain, Mohammad Mahdian, Evangelos Markakis, Amin Saberi, and Vijay V Vazirani. Greedy facility location algorithms analyzed using dual fitting with factor-revealing lp. *Journal of the ACM (JACM)*, 50(6):795–824, 2003.
- [19] David R. Karger and Maria Minkoff. Building steiner trees with incomplete global knowledge. In *FOCS*, pages 613–623. IEEE Computer Society, 2000.

- [20] Samir Khuller and Yoram J. Sussmann. The capacitated k-center problem. In *In Proceedings of the 4th Annual European Symposium on Algorithms, Lecture Notes in Computer Science 1136*, pages 152–166. Springer, 1996.
- [21] Samory Kpotufe. The curse of dimension in nonparametric regression. 2010.
- [22] Robert Krauthgamer and James R Lee. Navigating nets: simple algorithms for proximity search. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 798–807. Society for Industrial and Applied Mathematics, 2004.
- [23] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- [24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [25] Mu Li, David G Andersen, Alex J Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems*, pages 19–27, 2014.
- [26] Shanfei Li. An improved approximation algorithm for the hard uniform capacitated k-median problem. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2014, September 4-6, 2014, Barcelona, Spain*, pages 325–338, 2014.
- [27] Gaëlle Loosli, Stéphane Canu, and Léon Bottou. Training invariant support vector machines using selective sampling. *Large scale kernel machines*, pages 301–320, 2007.
- [28] Mohammad Mahdian and Martin Pál. Universal facility location. In *Algorithms-ESA 2003*, pages 409–421. Springer, 2003.
- [29] Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.
- [30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [31] Ruth Urner, Shai Shalev-Shwartz, and Shai Ben-David. Access to unlabeled data can speed up prediction time. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 641–648, 2011.
- [32] Ruth Urner, Sharon Wulff, and Shai Ben-David. Plal: Cluster-based active learning. In *Conference on Learning Theory*, pages 376–397, 2013.
- [33] Vladimir N. Vapnik and Leon Bottou. Local algorithms for pattern recognition and dependencies estimation. *Neural Computation*, 1993.
- [34] Yuchen Zhang, John Duchi, Michael Jordan, and Martin Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Neural Information Processing Systems*, 2013.
- [35] Yuchen Zhang, John C. Duchi, and Martin Wainwright. Communication-efficient algorithms for statistical optimization. In *Neural Information Processing Systems*, 2012.

# Appendix

## 6.1 Background

### Capacitated $k$ -center

The (uniform) capacitated  $k$ -center problem is to minimize the maximum distance between a cluster center and any point in its cluster subject to the constraint that the maximum size of a cluster is  $L$ . It is NP-Hard, so research has focused on finding approximation algorithms. Bar-Ilan et al [7] introduced the problem and presented the first constant factor polynomial time algorithm achieving a factor of 10, which was subsequently improved by Khuller et al [20]. It was a combinatorial algorithm that first guessed the optimal objective and a graph,  $G$  of all points with an edge between two points iff they are separated by less than the guess. They construct a set of monarchs: a maximal independent set in  $G^2$ , assign points close to the monarchs as their “empires”, move points around empires and open new centers are required to satisfy the capacity constraints, while increasing the objective in a bounded manner. This essentially means that a point cannot be assigned to a cluster arbitrarily far away from it. In [13], the capacitated  $k$ -center problem with non-uniform capacities is written as a feasible point of an integer linear program, and a procedure to round fractional solutions obtained from the LP relaxation is described. They construct a path-like tree structure with nearby vertices close in the original graph, and all vertices with fractional opening values at the leaves (which they call a “caterpillar” structure), transfer “openings” between distant vertices by transferring a limited number of clients to neighboring facilities through a chain (so as to increase the objective only in a bounded manner), and end up with all integral openings in polynomial time. Assignments are then made via a bipartite matching between points and enters as Hall’s marriage theorem can now be applied. The approximation factor is not explicitly computed, although it is mentioned to be “in the order of hundreds”.

[2] follows a similar procedure. They describe “tree instances” as generalizations of caterpillar structures of [13], and use a rounding procedure that is somewhat similar to the previous approach to get an approximation factor of 8. Further, for the special case of uniform capacities, they show a 6-approximation.

### Capacitated $k$ -median

$k$ -median with capacities is a notoriously difficult problem in clustering. It is much less understood than  $k$ -center with capacities, and uncapacitated  $k$ -median, both of which have constant factor approximations. Despite numerous attempts by various researchers, still there is no known constant factor approximation for capacitated  $k$ -median (even though there is no better lower bound for the problem than the one for uncapacitated  $k$ -median). As stated earlier, there is a well-known unbounded integrality gap for the standard LP even when violating the capacity or center constraints by a factor of  $2 - \epsilon$  [1].

Charikar et al. gave a 16-approximation when constraints are violated by a factor of 3 [11]. Byrka et al. improved this violation to  $2 + \epsilon$ , while maintaining an  $O(\frac{1}{\epsilon^2})$  approximation [10]. Recently, Li improved the latter to  $O(\frac{1}{\epsilon})$ , specifically, when constraints are violated by  $2 + \frac{2}{\alpha}$  for  $\alpha \geq 4$ , they give a  $6 + 10\alpha$  approximation [26]. These results are all for the *hard* capacitated  $k$ -median problem. In the *soft* capacities variant, we can open a point more than once to achieve more capacity, although each extra opening counts toward the budget of  $k$  centers. In hard capacities, each center can only be opened once. The hard capacitated version is more general, as each center can be replicated enough times so that the soft capacitated case reduces to the hard capacitated case. Therefore, we will only discuss the hard capacitated case.

All of the algorithms for capacitated  $k$ -median mentioned above share the same high-level idea but with different refinements in the algorithm and analysis. They are all LP rounding algorithms. They work by first using a monarch procedure to aggregate fractional center openings, where each demand is only moved a constant factor away. Each cluster must have at least  $\frac{1}{2}$  (or  $\frac{\alpha-1}{\alpha}$ ) total opening after this step. Then we must partition the fractional openings into star structures, and round the openings within each star.

### **$k$ -center with lower bounds**

Ene et al [15] describe a 4-approximation to  $k$ -center with lower bounds by constructing  $r$ -nets. They describe an efficient algorithm to this end. The reduction is specific to this objective and does not work with upper bounds on cluster sizes.

### **Universal and load balanced facility location**

In the facility location problem, we are given a set of demands and a set of possible locations for facilities. We should open facilities at some of these locations, and connect each demand point to an open facility so as to minimize the total cost of opening facilities and connecting demands to facilities. Capacitated facility location is a variant where each facility can supply only a limited amount of the commodity. This and other special cases are captured by the Universal Facility Location problem where the facility costs are general concave functions. Local search techniques [28] have been proposed and applied successfully. Also, LP rounding techniques suffer from unbounded integrality gap for capacitated facility location [28].

Load-balanced facility location [19], [17], is yet another variant where every open facility must cater to a minimum amount of demand. An unconstrained facility location problem with modified costs is constructed and solved. Every open facility that does not satisfy the capacity constraint is closed and the demand is rerouted to nearby centers. The modified problem is constructed so as to keep this increase in cost bounded.

## **7 Details from Section 2**

In this section, we provide the formal details for the bicriteria algorithm presented in Section 2.

The  $k$ -center LP is a little different from the  $k$ -median/means LP. As in prior work [2, 13, 20], we guess the optimal radius,  $t$ . Since there are a polynomial number of possibilities, we can try all of them to find the minimum possible  $t$  for which program 3 is feasible. Here is the LP for  $k$ -center.

$$\sum_{i \in V} x_{ij} = p, \quad \forall j \in V \quad (3a)$$

$$n\ell y_i \leq \sum_{j \in V} x_{ij} \leq nLy_i, \quad \forall i \in V \quad (3b)$$

$$\sum_{i \in V} y_i \leq k; \quad (3c)$$

$$0 \leq x_{ij} \leq y_i, \quad \forall i, j \in V \quad (3d)$$

$$x_{ij} = 0 \quad \text{if } d(i, j) > t. \quad (3e)$$

For  $k$ -median and  $k$ -means, let  $C_{LP}$  denote the objective value. For  $k$ -center,  $C_{LP}$  would be the smallest threshold  $t$  at which the LP is feasible, however we scale it as  $C_{LP} = tnp$  for consistency with the other objectives. For all  $j \in V$ , define the connection cost  $C_j$  as the average contribution of a point to the objective. For  $k$ -median and  $k$ -means, it is  $C_j = \frac{1}{p} \sum_{i \in V} c_{ij} x_{ij}$ . That is, for  $k$ -median, it is the average distance of a point to its fractional centers while for  $k$ -means, it is the average squared distance of a point to its fractional centers. For  $k$ -center,  $C_j$  is simply the threshold  $C_j = t$ . Therefore,  $C_{LP} = \sum_{j \in V} pC_j$  in all cases.

The notation is summarized in table 1.

**Step 2 details** Let  $\mathcal{M}$  be the set of monarchs, and for each  $u \in \mathcal{M}$ , denote  $\mathcal{E}_u$  as the empire of monarch  $u$ . Recall that the contribution of an assignment to the objective  $c_{ij}$  is  $d(i, j)$  for  $k$ -median,  $d(i, j)^2$  for  $k$ -means, and  $t$  for  $k$ -center. We also define a parameter  $\rho = 1$  for  $k$ -center,  $\rho = 2$  for  $k$ -median, and  $\rho = 4$  for  $k$ -means, for convenience.

Initially set  $\mathcal{M} = \emptyset$ . Order all points in nondecreasing order of  $C_i$ . For each point  $i$ , if  $\exists j \in \mathcal{M}$  such that  $c_{ij} \leq 2tC_i$ , continue. Else, set  $\mathcal{M} = \mathcal{M} \cup \{i\}$ . At the end of the for loop, assign each point  $i$  to cluster  $\mathcal{E}_u$  such that  $u$  is the closest point in  $\mathcal{M}$  to  $i$ . See Algorithm 3.

**Table 1:** Notation table

Symbol	Description	$k$ -median	$k$ -means	$k$ -center
$y_i$	Fractional opening at center $i$	-		
$x_{ij}$	Fractional assignment of point $j$ to center $i$	-		
$c_{ij}$	Cost of assigning $j$ to center $i$	$d(i, j)$	$d(i, j)^2$	$t$
$C_j$	Avg cost of assignment of point $j$ to all its centers	$\sum_i c_{ij} x_{ij} / p$		$= t$
$C_{LP}$	Cost of LP	$\sum_j p C_j$		
$\rho$	parameter for monarch procedure	2	4	1

<b>Input:</b> $V$ and fractional $(x, y)$	
<b>Output:</b> Set of monarchs, $\mathcal{M}$ , and empire $\mathcal{E}_j$ for each monarch $j \in \mathcal{M}$	
1	$\mathcal{M} \leftarrow \emptyset$
2	Order all points in non-decreasing order of $C_i$
3	// Identify Monarchs
4	<b>foreach</b> $i \in V$ <b>do</b>
5	<b>if</b> $\nexists j \in \mathcal{M}$ such that $c_{ij} \leq 2\rho C_i$ <b>then</b>
6	$\mathcal{M} \leftarrow \mathcal{M} \cup \{i\}$
7	// Assign Empires as Voronoi partitions around monarchs
8	<b>foreach</b> $j \in V$ <b>do</b>
9	Let $u \in \mathcal{M}$ be the closest monarch to $j$
10	$\mathcal{E}_u \leftarrow \mathcal{E}_u \cup \{j\}$

**Algorithm 3: Monarch procedure for coarse clustering:** Greedy algorithm to create monarchs and assign empires

Note that we can generalize Lemma 1 to hold for all three objectives. For Properties (1b) and (1c), we replace  $4C_j$  with  $2\rho C_j$  to generalize for  $k$ -means and  $k$ -center. For Property (1d), we show for  $k$ -center,  $\sum_{j \in \mathcal{E}_u} y_j \geq p$ , whereas  $k$ -means is the same as  $k$ -median ( $\geq \frac{p}{2}$ , not  $p$ ).

*Proof of Lemma 1.* The first three properties follow easily from construction (for the third property, recall we ordered the points at the start of the monarch procedure). Here is the proof of the final property, depending on the objective function.

For  $k$ -center and  $k$ -median, it is clear that for some  $u \in \mathcal{M}$ , if  $d(i, u) \leq \rho C_u$ , then  $i \in \mathcal{E}_u$  (from the triangle inequality and Property (1c)). For  $k$ -means, however: if  $d(i, u)^2 \leq 2C_u$ , then  $i \in \mathcal{E}_u$ . Note that the factor is  $\rho/2$  for  $k$ -means. This is because of the triangle inequality is a little different for squared distances.

To see why this is true for  $k$ -means, assume towards contradiction that  $\exists i \in V$ ,  $u, u' \in \mathcal{M}$ ,  $u \neq u'$  such that  $u \in \mathcal{E}_{u'}$  and  $d(i, u)^2 \leq 2C_u$ . Then  $d(i, u') \leq d(i, u)$  by construction. Therefore,  $d(u, u')^2 \leq (d(u, i) + d(i, u'))^2 \leq 4d(i, u)^2 \leq 8C_u$ , and we have reached a contradiction by Property (1c).

Now, to prove property (1d):

**$k$ -center** From the LP constraints, for every  $u$ ,  $\sum_{j \in V} x_{ju} = p$ . But  $x_{ju}$  is non-zero only if they are separated by at most  $t$ , the threshold. Combining this with the fact that if  $d(j, u) \leq C_u = t$ , then  $j \in \mathcal{E}_u$ , we get, for each  $u \in \mathcal{M}$ :

$$\sum_{j \in \mathcal{E}_u} y_j \geq \sum_{j \in \mathcal{E}_u} x_{ju} = p$$

**$k$ -median and  $k$ -means** Note that  $C_u$  is a weighted average of costs  $c_{iu}$  with weights  $x_{iu}/p$ , i.e.,  $C_u = \sum_i c_{iu} x_{iu}/p$ . By Markov's inequality,

$$\sum_{j: c_{ju} > 2C_u} \frac{x_{ju}}{p} < \frac{C_u}{2C_u} = \frac{1}{2}$$

Combining this with the fact that if  $c_{ju} \leq 2C_u$ , then  $j \in \mathcal{E}_u$  for both  $k$ -median and  $k$ -means, we get, for each  $u \in \mathcal{M}$ :

$$\sum_{j \in \mathcal{E}_u} y_j \geq \sum_{j: c_{ju} \leq 2C_u} y_j \geq \sum_{j: c_{ju} \leq 2C_u} x_{ju} \geq \frac{p}{2}.$$

□

**Step 3 Details** We show that the move operation does not violate any of the LP constraints except the constraint that  $y_i \leq 1$ . Should we require  $\delta \leq \min(y_a, 1 - y_b)$ , the constraint  $y_i \leq 1$  would not be violated. But to get a bicriteria approximation, we allow this violation. The amount by which the objective gets worse can then be bounded by the triangle inequality.

**Lemma 12.** *The operation Move does not violate any of the LP constraints except possibly the constraint  $y_i \leq 1$  and the threshold constraint 3e of  $k$ -center.*

*Proof.* To show that the Move operation satisfies all the LP constraints, first note that the only quantities that change are  $y_a, y_b, x_{au}, x_{bu}$ ,  $\forall u \in V$ . Further,  $x, y$  satisfy all the constraints of the LP. Using this,

- Constraint LP.1: For every  $u$ ,  $\sum_i x'_{iu} = \sum_i x_{iu} = p$ .
- Constraint LP.2 (1):

$$\begin{aligned} \sum_u x'_{au} &= \sum_u x_{au}(1 - \delta/y_a) \leq nLy_a(1 - \delta/y_a) = nLy'_a \\ \sum_u x'_{bu} &= \sum_u x_{bu} + \sum_u x_{au} \cdot \delta/y_a \\ &\leq nLy_b + nLy_a \cdot \delta/y_a = nLy'_b \end{aligned}$$

- Constraint LP.2 (2):

$$\begin{aligned}\sum_u x'_{au} &= \sum_u x_{au}(1 - \delta/y_a) \geq n\ell y_a(1 - \delta/y_a) = n\ell y'_a \\ \sum_u x'_{bu} &= \sum_u x_{bu} + \sum_u x_{au} \cdot \delta/y_a \\ &\geq n\ell y_b + n\ell y_a \cdot \delta/y_a = n\ell y'_b\end{aligned}$$

- Constraint LP.3:  $\sum_i y'_i = \sum_i y_i \leq k$

- Constraint LP.4 (1):

$$\begin{aligned}x'_{au} &= x_{au}(1 - \delta/y_a) \leq y_a(1 - \delta/y_a) = y'_a \\ x'_{bu} &= x_{bu} + x_{au} \cdot \delta/y_a \leq y_b + y_a \cdot \delta/y_a = y'_b.\end{aligned}$$

- Non-negative constraint: this is true since  $\delta \leq y_a$ .

□

See Algorithm 4 for the aggregation procedure.

<p><b>Input:</b> <math>V</math>, fractional <math>(x, y)</math>, empires <math>\{\mathcal{E}_j\}</math>  <b>Output:</b> updated <math>(x, y)</math></p> <pre> 1 <b>foreach</b> <math>\mathcal{E}_u</math> <b>do</b> 2   Define <math>Y_u = \sum_{i \in \mathcal{E}_u} y_i</math>, <math>z_u = \frac{Y_u}{\lfloor Y_u \rfloor}</math>. 3   <b>while</b> <math>\exists v</math> s.t. <math>y_v \neq z_u</math> <b>do</b> 4     Let <math>v</math> be the point farthest from <math>u</math> with nonzero <math>y_v</math>. 5     Let <math>v'</math> be the point closest to <math>j</math> with <math>y_{v'} \neq z_u</math>. 6     Move <math>\min\{y_v, z_u - y_{v'}\}</math> units of opening from <math>y_v</math> to <math>y_{v'}</math>.</pre>
---

**Algorithm 4: Aggregation procedure**

Note for  $k$ -center, the result corresponding to Lemma 3 is that  $d(i, j) \leq 5t$ , for all  $j \in V$  whose opening moved to  $i$ .

*Proof of Lemma 3.  $k$ -center.* Use the fact that all  $C_j = t$ , and  $x_{ij} > 0 \implies d(i, j) \leq t$  with property (1b) to get:

$$\begin{aligned}d(i, j) &\leq d(i, u) + d(u, i') + d(i', j) \\ &\leq 2C_i + 2C_{i'} + d(i', j) \leq 5t.\end{aligned}$$

**$k$ -means** The argument is similar to  $k$ -median, but with a bigger constant factor because of the squared triangle inequality.

$$\begin{aligned}
d(i, j)^2 &\leq (d(i, u) + d(u, i') + d(i', j))^2 \\
&\leq (2d(u, i') + d(i', j))^2 \\
&\leq 4d(u, i')^2 + d(i', j)^2 + 4d(u, i')d(i', j) \\
&\leq 4d(u, i')^2 + d(i', j)^2 + 4d(u, i')d(i', j) \\
&\quad + (2d(i', j) - d(u, i))^2 \\
&\leq 5d(u, i')^2 + 5d(i', j)^2 \\
&\leq 5d(j', i')^2 + 5d(i', j)^2 \\
&\leq 5(d(j', j) + d(j, i'))^2 + 5d(i', j)^2 \\
&\leq 5d(j', j)^2 + 10d(i', j)^2 + 10d(j', j)d(i', j) \\
&\leq 5d(j', j)^2 + 10d(i', j)^2 + 10d(j', j)d(i', j) \\
&\quad + 5(d(j', j) - d(i', j))^2 \\
&\leq 10d(j', j)^2 + 15d(i', j)^2 \\
&\leq 80C_j + 15d(i', j)^2.
\end{aligned}$$

□

**Step 4 details** Set  $\{i \mid y_i \neq 0\} = Y$ . We show details of the min cost flow network in Algorithm 5.

**Input:**  $V, (x, y), y$  are integral

**Output:** updated  $(x, y)$  with integral  $x$ 's and  $y$ 's

- 1 Create a flow graph  $G = (V', E)$  as follows.
- 2 Add each  $i \in V$  to  $V'$ , and give  $i$  supply  $p$ .
- 3 Add each  $i \in Y$  to  $V'$ , and give  $i$  demand  $n\ell$ .
- 4 Add a directed edge  $(i, j)$  for each  $i \in V, j \in Y$ , with capacity 2 and cost  $c_{ij}$  (for  $k$ -center, make the edge weight  $5t$  if  $d(i, j) \leq 5t$  and  $+\infty$  otherwise).
- 5 Add a sink vertex  $v$  to  $V'$ , with demand  $np - kn\ell$ .
- 6 Add a directed edge  $(i, v)$  for each  $i \in Y$ , with capacity  $\lceil \frac{p+2}{p}nL \rceil - n\ell$  and cost 0.
- 7 Run an min cost integral flow solver on  $G$ .
- 8 Update  $x$  by setting  $x_{ij}$  to 0, 1, or 2 based on the amount of flow going from  $i$  to  $j$ .

**Algorithm 5: Min cost flow procedure:** Set up flow problem to round  $x$ 's

**Lemma 13.** *There exists an integral assignment of the  $x'_{ij}$ 's such that  $\forall i, j \in V, x'_{ij} \leq 2$  and it can be found in polynomial time.*

*Proof.* See Algorithm 5 and Figure 2 for the details of the flow construction.

In this graph, there exists a feasible flow:  $\forall i, j \in V$ , send  $x'_{ij}$  units of flow along the edge from  $i$  to  $j$ , and send  $\sum_{j \in V} x_{ij}$  units of flow along the edge from  $i$  to  $v$ . Therefore, by the integral flow theorem, there exists a maximal integral flow which we can find in polynomial time. Also, by construction, this flow corresponds to an integral assignment of the  $x'_{ij}$ 's such that  $x'_{ij} \leq 2$ . □

*Proof of Theorem 4.*

**$k$ -center:** Recall that we defined  $C_{LP} = tnp$ , where  $t$  is the threshold for the  $k$ -center LP. From Lemma 3, when we reassign the demand of point  $j$  from  $i'$  to  $i$ ,  $d(i, j) \leq 5t$ . In other words, the  $y$ -rounded solution is



feasible at threshold  $5t$ . Then the  $k$ -center cost of the new  $y$ 's is  $np(5t) = 5C_{LP}$ . From Lemma 13, we can also round the  $x$ 's at no additional cost.

**$k$ -median:** From Property 3, when we reassign the demand of point  $j$  from  $i'$  to  $i$ ,  $d(i, j) \leq 3d(i', j) + 8C_j$ . Then we can bound the cost of the new assignments with respect to the original LP solution as follows.

$$\begin{aligned}
\sum_{i \in V} \sum_{j \in V} d(i, j) x'_{ij} &\leq \sum_{i \in V} \sum_{j \in V} (8C_j + 3d(i, j)) x_{ij} \\
&\leq \sum_{i \in V} \sum_{j \in V} 8C_j x_{ij} \\
&\quad + \sum_{i \in V} \sum_{j \in V} 3d(i, j) x_{ij} \\
&\leq \sum_{j \in V} 8C_j \sum_{i \in V} x_{ij} + 3C_{LP} \\
&\leq \sum_{j \in V} 8pC_j + 3C_{LP} \leq 11C_{LP}.
\end{aligned}$$

Then from Lemma 13, we get a solution of cost at most  $11C_{LP}$ , which also has integral  $x$ 's.

**$k$ -means:** The proof is similar to the  $k$ -median proof. From lemma 3, when we reassign the demand of point  $j$  from  $i'$  to  $i$ ,  $d(i, j)^2 \leq 15d(i', j)^2 + 80C_j$ . Then we can bound the cost of the new assignments with respect to the original LP solution as follows.

$$\begin{aligned}
\sum_{i \in V} \sum_{j \in V} d(i, j)^2 x'_{ij} &\leq \sum_{i \in V} \sum_{j \in V} (80C_j + 15d(i', j)^2) x_{ij} \\
&\leq \sum_{i \in V} \sum_{j \in V} 80C_j x_{ij} + \sum_{i \in V} \sum_{j \in V} 15d(i, j)^2 x_{ij} \\
&\leq \sum_{j \in V} 80C_j \sum_{i \in V} x_{ij} + 15C_{LP} \\
&\leq \sum_{j \in V} 80pC_j + 15C_{LP} \leq 95C_{LP}.
\end{aligned}$$

Then from Lemma 13, we get a solution of cost at most  $95C_{LP}$ , which also has integral  $x$ 's. □

See Algorithm 6 for the final algorithm.

**Input:**  $V$

**Output:** Integral  $(x, y)$  corresponding to bicriteria clustering solution

- 1 Run a solver for the LP relaxation for  $k$ -median,  $k$ -means, or  $k$ -center, output  $(x, y)$ .
- 2 Run Algorithm 3 with  $V$ ,  $(x, y)$ , output set of empires  $\{\mathcal{E}_j\}$ .
- 3 Run Algorithm 4 with  $V$ ,  $\{\mathcal{E}_j\}$ ,  $(x, y)$ , output updated  $(x, y)$ .
- 4 Run Algorithm 5 with  $V$ ,  $(x, y)$ , output updated  $(x, y)$ .

**Algorithm 6: Bicriteria approximation Algorithm for  $k$ -median,  $k$ -means, and  $k$ -center**

## 8 $k$ -center

In this section, we present a more complicated algorithm that is specific to  $k$ -center, which achieves a true approximation algorithm - the capacities are no longer violated.

## Approach

As in the previous section and in prior work [2, 13, 20], we start off by guessing the optimal distance  $t$ . Since there are a polynomial number of possibilities, it is still only polynomially expensive. We then construct the threshold graph  $G_t = (V, E_t)$ , with  $j$  being the set of all points, and  $(x, y) \in E_t$  iff  $d(x, y) \leq t$ .

A high-level overview of the rounding algorithm that follows is given in Algorithm 7.

**Connection to the previous section** The algorithm here is similar to the bicriteria algorithm presented previously. There are, however, two differences. Firstly, we work only with connected components of the threshold graph. This is necessary to circumvent the unbounded integrality gap of the LP [13]. Secondly, the rounding procedure of the  $y$ 's can now move opening across different empires. Since the threshold graph is connected, the distance between any two adjacent monarchs is bounded and turns out to exactly be thrice the threshold. This enables us to get a constant factor approximation without violating any constraints.

<p><b>Input:</b> <math>V</math>: the set of points, <math>k</math>: the number of clusters, <math>(\ell, L)</math>: min and max allowed cluster size</p> <p><b>Output:</b> A <math>k</math>-clustering of <math>V</math> respecting cluster size constraints, <math>p</math>: replication factor</p> <p><b>Procedure</b> <code>balanced-k-center</code>(<math>V, k, p, \ell, L</math>)</p> <div style="margin-left: 20px;"> <p><b>foreach</b> threshold <math>t</math> <b>do</b></p> <div style="margin-left: 20px;"> <p>Construct the threshold graph <math>G_t</math></p> <p><b>foreach</b> connected component <math>G^{(c)}</math> of <math>G_t</math> <b>do</b></p> <div style="margin-left: 20px;"> <p><b>foreach</b> <math>k'</math> in <math>1, \dots, k</math> <b>do</b></p> <div style="margin-left: 20px;"> <p>// Solve balanced <math>k'</math>-clustering on <math>G^{(c)}</math></p> <p>Solve <code>LPRound</code>(<math>G^{(c)}, k', p, \ell, L</math>)</p> </div> </div> <p>Find a solution for each <math>G^{(c)}</math> with <math>k_c</math> centers such that <math>\sum_c k_c = k</math> by linear search; call is <math>s</math></p> <p><b>if</b> no such a solution exists <b>then return</b> "No Solution Found"</p> <p><b>else return</b> solution <math>s</math></p> </div> </div> <p><b>Procedure</b> <code>LPRound</code>(<math>G, k, p, \ell, L</math>)</p> <div style="margin-left: 20px;"> <p><math>(x, y) \leftarrow</math> relaxed solution of LP in equation 4</p> <p><math>(x', y') \leftarrow</math> <code>yRound</code>(<math>G, x, y</math>)</p> <p>Round <math>x'</math> to get <math>x''</math> from theorem 20</p> <p><b>return</b> <math>(x'', y')</math></p> </div> <p><b>Procedure</b> <code>yRound</code>(<math>G, x, y</math>)</p> <div style="margin-left: 20px;"> <p>Construct coarse clustering to get a tree of clusters from algorithm 8</p> <p>Round clusters in a bottom up manner in the tree, moving mass around to nodes within a distance of 5 away (algorithm 9)</p> <p><b>return</b> rounded solution with integral <math>y</math></p> </div>
--

**Algorithm 7:** Algorithm overview

## The Algorithm

### Intuition

The approach is to guess the optimal threshold, construct the threshold graph at this threshold, write and round several LPs for each connected component of this graph for different values of  $k$ . The intuition behind why this works is that at the optimal threshold, each cluster is fully contained within a connected component (by definition of the threshold graph).

We the round the opening variables, but this time, open exactly  $k$  centers. Most of the work goes into rounding the openings, and showing that it is correct. Then, we simply round the assignments using a minimum cost flow again.

### Linear Program

As earlier, let  $y_i$  be an indicator variable to denote whether vertex  $i$  is a center, and  $x_{ij}$  be indicators for whether  $j$  belongs to the cluster centered at  $i$ . By convention,  $i$  is called a facility and  $j$  is called a client.

Consider the following LP relaxation for the IP for each connected component of  $G$ . Note that it is exactly the same as the one from the previous section, except it is described in terms of the threshold graph  $G$ . Let us call it  $\text{LP-}k\text{-center}(G)$ :

$$\sum_{i \in V} y_i = k \tag{4a}$$

$$x_{ij} \leq y_i \quad \forall i, j \in V \tag{4b}$$

$$\sum_{j:ij \in E} x_{ij} \leq nLy_i \quad \forall i \in V \tag{4c}$$

$$\sum_{j:ij \in E} x_{ij} \geq n\ell y_i \quad \forall i \in V \tag{4d}$$

$$\sum_{i:ij \in E} x_{ij} = p \quad \forall j \in V \tag{4e}$$

$$x_{ij} = 0 \quad \forall ij \notin E \tag{4f}$$

$$0 \leq x, y \leq 1 \tag{4g}$$

Once we have the threshold graph, for the purpose of  $k$ -center, all distances can now be measured in terms of the length of the shortest path in the threshold graph. Let  $d_G(i, j)$  represent the distance between  $i$  and  $j$  measured by the length of the shortest path between  $i$  and  $j$  in  $G$ .

### Connected Components

It is well known [13] that even without lower bounds and replication, the LP has unbounded integrality gap for general graphs. However, for connected components of the threshold graph, this is not the case.

To begin with, we show that it suffices to be able to do the LP rounding procedure for only connected threshold graphs, even in our generalization.

**Theorem 14.** *If there exists an algorithm that takes as input a connected graph  $G$ , capacities  $\ell, L$ , replication  $p$ , and  $k$  for which  $\text{LP-}k\text{-center}(G_t)$  is feasible, and computes a set of  $k$  centers to open and an assignment of every vertex  $j$  to  $p$  centers  $i$  such that  $d_G(i, j) \leq r$  satisfying the capacity constraints, then we can obtain a  $r$ -approximation algorithm to the balanced  $k$ -centers problem with  $p$ -replication.*

*Proof.* Let connected component  $i$  have  $k_i$  clusters. For each connected component, do a linear search on the range  $[1, \dots, k]$  to find values of  $k_i$  for which the problem is feasible. These feasible values will form a range, if size constraints are to be satisfied. To see why this is the case, note that if  $(x_1, y_1)$  and  $(x_2, y_2)$  are fractional solutions for  $k = k_1$  and  $k = k_2$  respectively, then  $((x_1 + x_2)/2, (y_1 + y_2)/2)$  is a valid fractional solution for  $k = (k_1 + k_2)/2$ .

Suppose the feasible values of  $k_i$  are  $m_i \leq k_i \leq M_i$ . If  $\sum_i m_i > k$  or  $\sum_i M_i < k$ , return NO (at this threshold  $t$ ). Otherwise, start with each  $k_i$  equal to  $m_i$ . Increase them one by one up to  $M_i$  until  $\sum_i k_i = k$ . This process takes polynomial time.  $\square$

From now on, the focus is entirely on a single connected component.

## Rounding $y$

Given an integer feasible point to the IP for each connected component, we can obtain the desired clustering. Hence, we must find a way to obtain an integer feasible point from any feasible point of **LP- $k$ -center**.

To round the  $y$ , we follow the approach of An et al.

citeckc3. The basic idea is to create a coarse clustering of vertices, and have the cluster centers form a tree. The radius of each cluster will be at most 2, and the length of any edge in the tree will exactly be three, by construction.

Now, to round the  $y$ , we first start from the leaves of the tree, moving opening around in each coarse cluster such that at most one node (which we pick to be the center, also called the monarch). In subsequent steps, this fractional opening is passed to the parent cluster, where the same process happens. The key to getting a constant factor approximation is to ensure that fractional openings that transferred from a child cluster to a parent cluster are not propagated further. Note that the bicriteria algorithm did not move opening from one coarse cluster (empire) to another because we didn't have an upper bound of the cost incurred by making this shift.

**Preliminaries.** We start with some definitions.

**Definition 2** ( $\delta$ -feasible solution [13]). *A solution  $(x, y)$  feasible on  $G^\delta$ , the graph obtained by connecting all nodes within  $\delta$  hops away from each other.*

Next, we introduce the notion of a distance- $r$  shift. Intuitively, a distance- $r$  shift is a series of movements of openings, none of which traverses a distance more than  $r$  in the threshold graph. Note that the definition is similar to what is used in An et al.

citeckc3.

**Definition 3** (Distance- $r$  shift ). *Given a graph  $G = (V, E)$  and  $y, y' \in \mathbb{R}_{\geq 0}^{|V|}$ ,  $y'$  is a distance- $r$  shift of  $y$  if  $y'$  can be obtained from  $y$  via a series of disjoint movements of the form “Move  $\delta$  from  $i$  to  $i'$ ” where  $\delta \leq \min(y_i, 1 - y_{i'})$  and every  $i$  and  $i'$  are at most a distance  $r$  apart in the threshold graph  $G$ . Further, if all  $y'$  are zero or one, it is called an integral distance- $r$  shift.*

Note that, by the definition of a distance- $r$  shift, each unit of  $y$  moves only once and if it moves more than once, all the movements are put together as a single, big movement, and this distance still does not exceed  $r$ .

**Lemma 15** (Realizing distance- $r$  shift). *For every distance- $r$  shift  $y'$  of  $y$  such that  $0 \leq y'_i \leq 1 \forall i \in V$ , we can find  $x'$  in polynomial time such that  $(x', y')$  is  $(r+1)$ -feasible.*

*Proof.* We can use the Move operation described earlier and in Cygan et al. [13] to change the corresponding  $x$  for each such a movement to ensure that the resulting  $(x', y')$  are  $(r+1)$ -feasible. The additional restriction  $\delta \leq 1 - y_{i'}$  ensures that  $y \leq 1$ . Since each unit of  $y$  moves only once, all the movements put together will also lead a solution feasible in  $G^{r+1}$ , i.e. we get a  $(r+1)$ -feasible solution. □

From here on, we assume that  $x_{ij}, x_{i'j}$  are adjusted as described above for every movement between  $i$  and  $i'$ .

The algorithm to round  $y$  [2] proceeds in two phases. In the first phase, we cluster points into a tree of coarse clusters (monarchs) such that nearby clusters are connected using the monarch procedure of Khuller et al [20]. In the second phase, fractional opening are aggregated to get an integral distance-5 shift.

**Monarch Procedure.** The monarch procedure presented a little differently but is very similar to the monarch procedure presented earlier. Since the threshold graph is connected, we can get guarantees on how big the distance between two monarchs is.

<p><b>Input:</b> <math>G = (V, E)</math></p> <p><b>Output:</b> Tree of monarchs, <math>T = (\mathcal{M}, E')</math>, and empires for each monarch</p> <pre> 1 Marked <math>\leftarrow \emptyset</math> 2 <b>foreach</b> <math>j \in V</math> <b>do</b> 3   <math>\quad</math> initialize <math>\text{ChildMonarchs}(j)</math> and <math>\text{Dependents}(j)</math> to <math>\emptyset</math> 4 Pick any vertex <math>u</math> and make it a monarch 5 <math>\mathcal{E}_u \leftarrow N^+(u)</math>; Initialize <math>T</math> to be a singleton node <math>u</math> 6 Marked <math>\leftarrow \text{Marked} \cup \mathcal{E}_u</math> 7 <b>while</b> <math>\exists w \in (V \setminus \text{Marked})</math> such that <math>d_G(w, \text{Marked}) \geq 2</math> <b>do</b> 8   <math>\quad</math> Let <math>u \in (V \setminus \text{Marked})</math> and <math>v \in \text{Marked}</math> such that <math>d_G(u, v) = 2</math> 9   <math>\quad</math> Make <math>u</math> a monarch and assign its empire to be <math>\mathcal{E}_u \leftarrow N^+(u)</math> 10  <math>\quad</math> Marked <math>\leftarrow \text{Marked} \cup \mathcal{E}_u</math> 11  <math>\quad</math> Make <math>u</math> a child of <math>m(v)</math> in <math>T</math> 12  <math>\quad</math> ChildMonarchs(<math>v</math>) <math>\leftarrow \text{ChildMonarchs}(v) \cup \{u\}</math> 13 <b>foreach</b> <math>v \in (V \setminus \text{Marked})</math> <b>do</b> 14   <math>\quad</math> Let <math>u \in \text{Marked}</math> be such that <math>d_G(u, v) = 1</math> 15   <math>\quad</math> Dependents(<math>u</math>) <math>\leftarrow \text{Dependents}(u) \cup \{v\}</math> 16   <math>\quad</math> <math>\mathcal{E}_{m(u)} \leftarrow \mathcal{E}_{m(u)} \cup \{v\}</math> </pre>
--

**Algorithm 8:** Monarch Procedure: Algorithm to construct tree of monarchs and assign empires

Algorithm 8 describes the first phase where we construct a tree of monarchs and assign empires to each monarch. Let  $\mathcal{M}$  be the set of all monarchs. For some monarch,  $u \in \mathcal{M}$ , let  $\mathcal{E}_u$  denote its empire. For each vertex  $i$ , let  $m(i)$  denote the monarch  $u$  to whose empire  $\mathcal{E}_u$ ,  $i$  belongs.

The guarantees now translate to the following (Lemma 16):

- Empires partition the point set.
- The empire includes *all* immediate neighbors of a monarch and additionally, some other nodes of distance two (called dependents).
- Adjacent monarchs are exactly distance 3 from each other.

**Lemma 16.** *Algorithm 8, the monarch procedure is well-defined and its output satisfies the following:*

- $\mathcal{E}_u \cap \mathcal{E}_{u'} = \emptyset$ .
- $\forall u \in \mathcal{M} : \mathcal{E}_u = N^+(u) \cup (\bigcup_{j \in N^+(u)} \text{Dependents}(j))$ .
- The distance between a monarch and any node in its empire is at most 2.
- Distance between any two monarchs adjacent in  $T$  is exactly 3.
- If  $\text{ChildMonarchs}(j) \neq \emptyset$  or  $\text{Dependents}(j) \neq \emptyset$ , then  $j$  is at distance one from some monarch.

*Proof.* Note that the whole graph is connected and  $V \neq \emptyset$ . For the while loop, if there exists  $w$  such that  $d_G(w, \text{Marked}) \geq 2$ , there exists  $u$  such that  $d_G(u, \text{Marked}) = 2$  because the graph is connected. By the end of the while loop, there are no vertices at a distance 2 or more from **Marked**. Hence, vertices not in **Marked**, if any, should be at a distance 1 from **Marked**. Thus, the algorithm is well defined.

Each time a new monarch  $u$  is created,  $N^+(u)$  is added to its empire. This shows the first statement. The only other vertices added to any empire are the dependents in the foreach loop. Each dependent  $j$  is directly connected to  $i$ , a marked vertex. Hence,  $i$  has to be a neighbor of a monarch. If  $i$  were a monarch,  $j$  would have been marked in the while loop. Thus,  $d_G(j, m(i)) = 2$ .

If the first statement of the while loop,  $v$  is a marked vertex, and has to be a neighbor of some monarch  $m(v)$ . New monarch  $u$  is chosen such that  $d_G(u, v) = 2$ . The parent monarch of  $u$  is  $m(v)$  and  $d_G(u, m(v)) = d_G(u, v) + d_G(v, m(v)) = 3$ . □

**Initial Aggregation.** Now, we shall turn to the rounding algorithm of An et al [2]. The algorithm begins with changing  $y_u$  of every monarch  $u \in \mathcal{M}$  to 1. Call this the initial aggregation. It requires transfer of at most distance one because the neighbors of the monarchs has enough opening.

**Lemma 17.** *The initial aggregation can be implemented by a distance-1 shift.*

*Proof.* For every vertex  $u \in V$ , we have  $\sum_{j \in N(u)} y_j \geq \sum_{j \in N(u)} x_{uj} = p \geq 1$ . Hence, there is enough  $y$ -mass within a distance of one from  $u$ . The actual transfer can happen by letting  $\delta = \min(1 - y_u, y_j)$  for some neighbor  $j$  of  $u$  and then transferring  $\delta$  from  $j$  to  $u$ . That is,  $y_j = y_j - \delta$  and  $y_u = y_u + \delta$ . □

**Rounding.** The rounding procedure now proceeds in a bottom-up manner on the tree of monarchs, rounding all  $y$  using movements of distance 5 or smaller. After rounding the leaf empires, all fractional opening, if any is at the monarch. For internal empires, the centers of child monarch (remnants of previous rounding steps) and dependents are first rounded. Then the neighbors of the monarch are rounded to leave the entire cluster integral except the monarch. The two step procedure is adopted so that the opening propagated from this monarch to its parent originates entirely from the 1-neighborhood of the monarch.

Formally, at the end of each run of round on  $u \in \mathcal{M}$ , all the vertices of the set  $I_u$  are integral, where  $I_u := (\mathcal{E}_u \setminus u) \cup (\bigcup_{j \in N(u)} \text{ChildMonarchs}(j))$ .

The rounding procedure is described in detail in Algorithm 9. The following lemma states and proves that algorithm 9 rounds all points and doesn't move opening very far.

**Lemma 18** (Adaptation of Lemma 19 of An et al [2]). *Let  $I_u := (\mathcal{E}_u \setminus u) \cup (\bigcup_{j \in N(u)} \text{ChildMonarchs}(j))$ .*

- *Round( $u$ ) makes the vertices of  $I_u$  integral with a set of opening movements within  $I_u \cup \{u\}$ .*
- *This happens with no incoming movements to the monarch  $u$  after the initial aggregation.*
- *The maximum distance of these movements is five, taking the initial aggregation into account.*

*Proof. Integrality.* From lemma 16, it can be seen that  $X_j, j \in N(u)$  above form a partition of  $I_u$ . Hence, it suffices to verify that each node of every  $X_j$  is integral.

At the end of line 8, the total non-integral opening in  $X_j$  is  $y(X_j) - \lfloor y(X_j) \rfloor$ , and is hence smaller than one. Line 9 moves all these fractional openings to  $j$ . By now, all openings of  $X_j \setminus \{j\}$  are integral.

Now,  $F$  is the set of all non-integral  $j \in N(u)$ . So, by the end of line 13, the total non-integral opening in  $N(u)$  (and hence in all of  $I_u$ ) is  $y(F \setminus W_F) = y(F) - \lfloor y(F) \rfloor$ , and is again smaller than one. If this is zero, we are done.

Otherwise, we choose a node  $w^*$ , shift this amount to  $w^*$  in line 17. To make this integral, this operations also transfers the *remaining amount*, i.e.  $1 - y(F \setminus W_F)$  from the monarch  $u$ . If this happens, the monarch  $u$ 's opening is no longer integral, but  $I_u$ 's is.

This shows the first bullet. For the second one, notice that after the initial aggregation, this last operation is the only one involving the monarch  $u$  and hence, there are no other incoming movements into  $u$ .

**Distance.** In the first set of transfers in line 8 the distance of the transfer is at most 4. This is because dependents are a distance one away from  $j$  and child monarchs are at a distance two away. The maximum distance is when the transfer happens from one child monarch to another, and this distance is 4 (recall that there are no incoming movements into monarchs).

The transfers in line 9 moves openings from a child monarch or a dependent to  $j$ . The distances are 2 and 1 respectively. Accounting for the initial aggregation, this is at most 3.

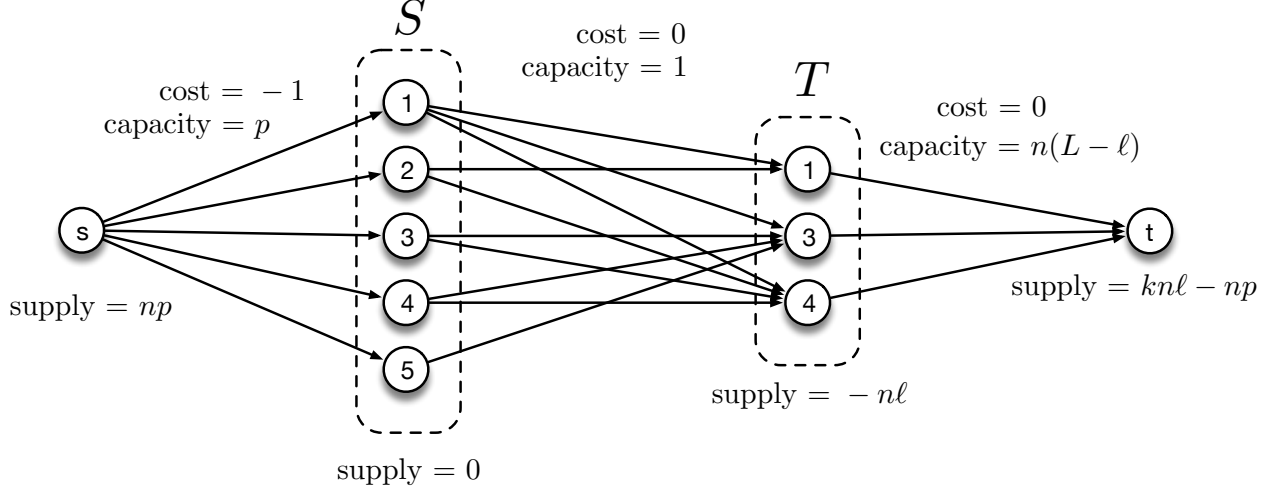
**Input:** Tree of monarchs,  $T$ , and empires for each monarch after the initial aggregation  
**Output:**  $y'$ , an integral distance-5 shift of  $y$

```

1 Procedure Round(Monarch  $u$ )
2   //Recursive call
3   foreach child  $w$  of  $u$  in  $T$  do Round( $w$ )
4   //Phase 1
5   foreach  $j \in N(u)$  do
6      $X_j \leftarrow \{j\} \cup \text{ChildMonarchs}(j) \cup \text{Dependents}(j)$ 
7      $W_j \leftarrow \{\lfloor y(X_j) \rfloor \text{ nodes from } X_j\}$ ; (Avoid picking  $j$  if possible)
8     LocalRound( $W_j, X_j, \emptyset$ )
9     LocalRound( $\{j\}, X_j \setminus W_j, \emptyset$ )
10  //Phase 2
11   $F = \{j | j \in N(u) \text{ and } 0 < y_j < 1\}$ 
12   $W_F \leftarrow \{\text{any } \lfloor y(F) \rfloor \text{ nodes from } F\}$ 
13  LocalRound( $W_F, F, \emptyset$ )
14  //Residual
15  if  $y(F \setminus W_F) > 0$  then
16    Choose  $w^* \in F \setminus W_F$ 
17    LocalRound( $\{w^*\}, F \setminus W_F, u$ )
18 Procedure LocalRound( $V_1, V_2, V_3$ )
19   while  $\exists i \in V_1$  such that  $y_i < 1$  do
20     Choose a vertex  $w$  with non-zero opening from  $V_2 \setminus V_1$ 
21     if there exists none, choose  $j$  from  $V_3 \setminus V_1$ 
22      $\delta \leftarrow \min(1 - y_i, y_j)$ 
23     Move  $\delta$  from  $j$  to  $i$ 

```

**Algorithm 9:** Algorithm to round  $y$



**Figure 8:** Minimum cost flow network to round  $x$ 's. Each node in a group has the same supply, which is indicated below. The cost and capacity of each edge is indicated above.

The rounding on line 13 moves openings between neighbors of the monarch, i.e. from some  $j$  to  $j'$  where  $j, j' \in N(u)$ . So, the distance between  $j$  and  $j'$  is at most 2. From the preceding transfers, the openings at  $j$  moved a distance of at most three to get there, and thus, we conclude that openings have moved at most a distance of 5 so far.

The first step of rounding on line 17 moves openings from some  $j$  to  $w^*$ , where  $j, w^* \in N(u)$ . As above, the maximum distance in this case is 5. The second step of rounding on line 17 moves opening from the monarch  $u$  to its neighbor  $w^*$ . This distance is one, and after accounting for the initial aggregation, is 2.

From this, we see that the maximum distance any opening has to move is 5.  $\square$

The algorithms, their properties in conjunction with lemma 15 leads to the following theorem, which also summarizes this subsection.

**Theorem 19.** *There exists a polynomial time algorithm to find a 6-feasible solution with all  $y$  integral.*

### Rounding $x$

Once we have integral  $y$ , rounding the  $x$  is fairly straight-forward, without making the approximation factor any worse. Exactly the same procedure used in bicriteria algorithms works here too. But, we can have an easier construction since for  $k$ -center since we can use distances in the threshold graph instead.

If there were no lower bounds on cluster sizes and no replication, this phase can be done by computing the a matching between points and centers to fix the cluster assignments ( $x$ ) [2]. Once we introduce replication, the assignments can be fixed by using maximum flows, which are generalization of matchings. Further generalizations are needed when we introduce lower bounds. Minimum cost flows are general enough to handle this extension.

**Theorem 20.** *There exists a polynomial time algorithm that given a  $\delta$ -feasible solution  $(x, y)$  with all  $y$  integral, finds a  $\delta$ -feasible solution  $(x', y)$  with all  $x'$  integral.*

*Proof.* We shall use a minimum cost flow network to this. Consider a directed bipartite graph  $(S, T, E')$ , where  $S = V$  and  $T = \{i : y_i = 1\}$  and  $j \rightarrow i \in E'$  iff  $x_{ij} > 0$ . Add a dummy vertex  $s$ , with edges to every vertex in  $S$ , and  $t$  with edges from every vertex in  $T$ . In this network, let every edge of the bipartite graph have capacity 1. Further, all the  $s \rightarrow S$  edges have capacity  $p$ .  $s$  supplies a flow of  $np$  units, while each  $u \in T$  has a demand of  $l$  units. To ensure no excess demand or supply,  $t$  has a demand of  $np - kl$ . All the  $t \rightarrow T$  edges have a capacity of  $(L - \ell)$ .



All the  $s \rightarrow S$  edges have a cost of  $-1$  and every other edge has a cost of zero. See figure 8.

Clearly, a feasible assignment  $(x, y)$  to  $\text{LP-}k\text{-center}(G^\delta)$  with integral  $y$  is a feasible flow in this network. In fact, it is a minimum-cost flow in this network. This can be verified by the absence of negative cost cycles in the residual graph (because all negative cost edges are at full capacities).

Since, the edge capacities are all integers, there exists a minimum cost integral flow by the Integral Flow Theorem. This flow can be used to fix the cluster assignments.  $\square$

Piecing together theorems 19 and 20, we have the following theorem:

**Theorem 21.** *Given an instance of the  $k$ -centers problem with  $p$ -replication and for a connected graph  $G$ , and a fractional feasible solution to  $\text{LP-}k\text{-center}(G)$ , there exists a polynomial time algorithm to obtain a 6-feasible integral solution. That is, for every  $i, j$  such that  $x_{ij} \neq 0$ , we have  $d_G(i, j) \leq 6$ .*

## 9 Proofs from Section 3

**Lemma 22.** *The cost of the objective of the star graph of size  $\geq 2l + 1$  strictly increases in  $k$ , for  $nl \geq 3$ .*

*Proof.* Let the size of the star graph be  $n$ . Clearly, the optimal center for  $k = 1$  is  $c$ . Then  $\text{OPT}_1 = n - 1$ . Then for  $k = 2$ , we must choose another center  $p$  that is not  $c$ .  $p$  is distance 2 to all points other than  $c$ , so the optimal clustering is for  $p$ 's cluster to have the minimum of  $nl$  points, and  $c$ 's cluster has the rest. Therefore,  $\text{OPT}_2 = n + nl - 2$ .

This process continues; every time a new center is added, the new center pays 0 instead of 1, but  $nl - 1$  new points must pay 2 instead of 1. This increases the objective by  $nl - 2$ . As long as  $nl \geq 3$ , this ensures the objective function is strictly increasing in  $k$ .  $\square$

**Lemma 23.** *For all  $k'$ , there exists a clustering instance in which the objective function has a local minimum at  $k'$ , even for soft capacities.*

*Proof.* Given  $l \geq 3$ , we create a clustering instance as follows. Define  $k'$  sets of points  $G_1, \dots, G_{k'}$ , each of size  $2nl - 1$ . For any two points in some  $G_i$ , set their distance to 0. For any two points in different sets, set their distance to 1.

Then for  $1 \leq k \leq k'$ , the objective value is equal to  $(k' - k)(2nl - 1)$ , since we can put  $k$  centers into  $k$  distinct groups, but  $(k' - k)$  groups will not have centers, incurring cost  $2nl - 1$ . When  $k > k'$ , we cannot put each center in a distinct group, so there is some group  $G_i$  with two centers. Since  $|G_i| = 2nl - 1$ , the two centers cannot satisfy the capacity constraint with points only from  $G_i$ , so the objective value increases.  $\square$

*proof of Lemma 5.* Consider the graph in Figure 4, and let  $nl = 21$ . We claim that there exist valid clusterings using only length 1 edges for  $k = 2$  and  $k = 4$ , but not  $k = 3$ . For  $k = 2$ , let the centers be  $y_1$  and  $y_2$ . Then  $y_1$  grabs the 20 red points (and itself) to hit the capacity of 21.  $y_2$  grabs all the rest of the points. For  $k = 4$ , let the centers be  $x_1, x_2, x_3, x_4$ . Then each have edges to 20 middle points, non-overlapping, and WLOG  $x_1$  grabs  $y_1$  and  $y_2$ .

Now consider  $k = 3$ . The crucial property is that by construction,  $y_1$  and any  $x_i$  cannot simultaneously be centers and each satisfy the capacity to distance 1 points. This is because all  $x_i$  and  $y_1$  are at minimum capacity, but  $y_1$ 's neighbors overlap with neighbors from each  $x_i$ . So we cannot just take the centers from  $k = 2$  and add a center from  $k = 4$ . The rest of the proof is checking that no other case works.

Case 1: the set of centers includes a point  $p$  not in  $\{x_1, x_2, x_3, x_4, y_1, y_2\}$ . The rest of the points are only distance 1 from exactly two points, so  $p$  cannot hit the lower bound of 21 using only distance 1 assignments.

Case 2: the set of centers is a subset of  $\{x_1, x_2, x_3, x_4\}$ . Then there are clearly 20 points which are not distance 1 from the three centers.

Case 3: the set of centers includes both  $y_1$  and  $y_2$ . Then we need to pick one more center,  $x_i$ .  $x_i$  is distance 1 from 20 middle points, plus  $\{x_1, x_2, x_3, x_4, y_1, y_2\}$ , so 26 total.  $y_1$  is also distance 1 from 20 middle points and  $\{x_1, x_2, x_3, x_4, y_1, y_2\}$ .  $y_1$  and  $x_i$  share exactly 5 neighbors from the middle points,

plus  $\{x_1, x_2, x_3, x_4, y_1, y_2\}$  as neighbors. Then the union of points that  $x_i$  and  $y_1$  are distance 1 from, is  $26 + 26 - 11 = 41$ , which implies that  $x_i$  and  $y_1$  cannot simultaneously reach the lower bound of 21 with only distance 1 points.

Case 4: the set of centers does not include  $x_i$  nor  $y_j$ . By construction, for each pair  $x_i$  and  $y_j$ , there exists some middle points which are only distance 1 from  $x_i$  and  $y_j$ .

These cases are exhaustive, so we conclude  $\mathcal{OPT}_3$  must be strictly larger than  $\mathcal{OPT}_2$  and  $\mathcal{OPT}_4$  (no matter what objective we use).  $\square$

*proof of Theorem 6. Setup.* Set  $k_{\min} = 10 \cdot m$ , and  $k_{\max} = 12m$ . Define  $K_{\text{good}} = \{k \mid k_{\min} \leq k \leq k_{\max} \text{ and } 2 \mid k\}$ . Similarly, let  $K_{\text{bad}} = \{k \mid k_{\min} \leq k \leq k_{\max} \text{ and } 2 \nmid k\}$ . Note  $|K_{\text{bad}}| = m$  and  $|K_{\text{good}}| = m + 1$ . For all  $k \in K_{\text{good}}$ , define  $X_k = \{x_1^{(k)}, \dots, x_{k'}^{(k)}\}$ . Let  $X = \bigcup_k X_k$ .

Define  $G = (V, E)$ ,  $V = X \cup Y$ ,  $X \cap Y = \emptyset$ . Just like in the last proof, the edges later correspond to a distance of 1, and all other distances are 2. We will construct  $Y$  and  $E$  such that for all  $k \in K_{\text{good}}$ , all the neighbors of  $X_k$  form a partition of  $Y$ , i.e.  $\forall k \in K_{\text{good}}, \bigcup_i N(x_i^{(k)}) = Y$  and  $N(x_i^{(k)}) \cap N(x_j^{(k)}) = \emptyset$  for all  $i \neq j$ . So taking  $X_k$  as the centers corresponds to a  $k$ -clustering in which all points are distance 1 from their center. We will also show that for all  $k \in K_{\text{bad}}$ , it is not possible to find a valid set of centers for which every point has an edge to its center, unless the capacities are violated. This implies that all  $m$  points in  $K_{\text{bad}}$  are local maxima.

For all  $k \in K_{\text{good}}$ ,  $X_{k'}$  will have exactly  $\frac{k_{\max}}{k'}l$  edges in  $Y$ . Thus, set  $n\ell = \prod_{k \in K_{\text{good}}} k$  to make all of these values integral. Note that some points (those in  $X_{k_{\max}}$ ) have exactly  $n\ell$  edges, and all points have  $\leq \frac{6}{5}n\ell$  edges (which is tight for the points in  $X_{k_{\min}}$ ).

Now we define the main property which drives the proof. We say  $x_{i_1}^{(j_1)}$  *overlaps* with  $x_{i_2}^{(j_2)}$  if  $N(x_{i_1}^{(j_1)}) \cup N(x_{i_2}^{(j_2)}) > \frac{2}{5}n\ell$ . Note this immediately implies it is not possible to include them in the same set of centers such that each point has an edge to its center, since  $N(x_{i_1}^{(j_1)}) \cup N(x_{i_2}^{(j_2)}) \leq N(x_{i_1}^{(j_1)}) + N(x_{i_2}^{(j_2)}) - N(x_{i_1}^{(j_1)}) \cap N(x_{i_2}^{(j_2)}) < 2 \cdot \frac{6}{5}n\ell - \frac{2}{5}n\ell = 2n\ell$ .

**Outline.** We will construct  $Y$  in three phases. First, we add edges to ensure that for all  $x_{i_1}^{(j_1)}$ , for all  $j_2 \neq j_1$ , there exists an  $i_2$  such that  $x_{i_1}^{(j_1)}$  overlaps with  $x_{i_2}^{(j_2)}$ . It follows that if we are trying to construct a set of centers from  $X$  for  $k' \in K_{\text{bad}}$ , we will not be able to use any complete  $X_{k'}$  as a subset. These are called the *backbone* edges.

The next phase is to add enough edges among points in different  $X_k$ 's so that no subset of  $X$  (other than the  $X_{k'}$ 's) is a complete partition of  $Y$ . We will accomplish this by adding a bunch of points to  $Y$  shared by various  $x \in X$ , so that each  $x$  has edges to  $k_{\max}$  points in  $Y$ . These are called the *dispersion* edges.

The final phase is merely to add edges so that all points reach their assigned capacity. We do this arbitrarily. These are called the *filler* edges.

Note whenever we add a point to  $Y$ , for all  $k \in K_{\text{good}}$ , we need to add an edge to exactly one  $x \in X_k$ , which will ensure that all  $X_k$ 's form a partition of  $Y$ .

**Phase 1: Backbone edges.** Recall that for  $k, k' \in K_{\text{good}}$ , we want  $\forall i, \exists j$  such that  $x_i^{(k)}$  overlaps with  $x_j^{(k')}$ . Since  $k_{\max} = \frac{6}{5}k_{\min}$ , some  $x$ 's will be forced to overlap with two points from the same  $X_k$ . However, we can ensure no point overlaps with three points from the same  $X_k$ .

We satisfy all overlappings naturally by creating  $k_{\min}$  components,  $CC_1$  to  $CC_{k_{\min}}$ . Each component  $CC_i$  contains point  $x_i^{(k_{\min})}$ . The rest of the sets  $X_k$  are divided so that one or two points are in each component, as shown in Figure 5 in Section 3. Formally, in component  $CC_i$ , sets  $X_{k_{\min}}$  to  $X_{k_{\min} + \lceil \frac{i}{2} \rceil}$  have one point in the component, and all other sets have two points in the component.

For each component  $CC_i$ , we add  $\frac{4}{5}n\ell$  points to  $Y$ , split into two groups of  $\frac{2}{5}n\ell$ . The points from sets  $X_{k_{\min} + \lceil \frac{i}{2} \rceil}$  have edges to all  $\frac{4}{5}n\ell$  points, and the points from the rest of the sets (since there are two from each set) have edges to one group of  $\frac{2}{5}n\ell$  points. Therefore, for all  $k, k' \in K_{\text{good}}$ , each point  $x \in X_k$  belongs to some component  $CC_i$ , and overlaps with some  $x' \in X_{k'}$ , so all of the overlapping requirements are satisfied (only using points within the same component).

This completes phase 1. Each point in  $X$  had at most  $\frac{4}{5}n\ell$  edges added, so every point can still take at least  $\frac{n\ell}{5}$  more edges in subsequent phases.

**Phase 2: Dispersion edges.** Now we want to add points to  $Y$  to ensure that no set of at most  $k_{max}$  points from  $X$  create a partition of  $Y$ , except sets that completely contain some  $X_k$ .

We have a simple way of achieving this. For every  $(x_1, x_2, \dots, x_{m+1}) \in X_{k_{min}} \times X_{k_{min}+2} \times \dots \times X_{k_{max}}$ , add one point to  $Y$  with edges to  $x_1, x_2, \dots, x_{m+1}$ . Then we have added  $\prod_{k \in K_{good}} k$  total points to  $Y$  in this phase.

This completes phase 2.

**Phase 3: Filler edges.** The final step is just to fill in the leftover points, since we want every point  $x_i^{(k)}$  to have  $\frac{k_{min}}{k}l$  points total. All of the mechanisms for the proof have been set up in phases 1 and 2, so these final points can be arbitrary.

We greedily assign points. Give each point  $x_i^{(k)} \in X$  a number  $t_{x_i^{(k)}} = \frac{k_{min}}{k}n\ell - N(x_i^{(k)})$ , i.e., the number of extra points it needs. Take the point  $x \in X_k$  with the minimum  $t$ , and create  $t$  points in  $Y$  with  $x$ . For each layer other than  $X_k$ , add edges to the point with the smallest number. Continue this process until  $t = 0$  for all points.

**Final Proof** Now we are ready to prove that  $G$  has  $m$  local maxima. By construction, for all  $k \in K_{good}$ ,  $X_k$  is a set of centers which satisfy the capacity constraints, and every point has an edge to its center. Now, consider a set  $C$  of centers of size  $k' \in K_{bad}$ . We show in every case,  $C$  cannot satisfy the capacity constraints with all points having edges to their centers.

Case 1:  $C$  contains a point  $y \in Y$ .  $y$  only has  $m$  edges, which is much smaller than  $n\ell$ .

Case 2: There exists  $k \in K_{good}$  such that  $X_k \subseteq C$ . Then since  $|C| \notin K_{good}$ ,  $\exists x \in C \setminus X_k$ . By construction, there exists  $x_i^{(k)} \in X_k$  such that  $x$  and  $x_i^{(k)}$  are overlapping. Therefore, both centers cannot satisfy the capacity constraints with points they have an edge to.

Case 3: For all  $k \in K_{good}$ , there exists  $x \in X_k$  such that  $x \notin C$ . Take the set of all of these points,  $x_1, x_2, \dots, x_{m+1}$ . By construction, there is a point  $y \in Y$  with edges to only these points. Therefore,  $y$  will not have an edge to its center in this case.

This completes the proof. □

## 10 Proofs from Section 4

We begin by introducing some notation that will be useful in the following proofs. At the population level, for any lower bound  $\ell$  and upper bound  $L$  on the cluster capacities, we denote the set of cluster assignments that satisfy the capacity constraints by

$$F(\ell, L) = \left\{ f : \mathcal{X} \rightarrow \binom{[k]}{p} : \mathbb{P}_{x \sim \mu} (i \in f(x)) \in [\ell, L] \forall i \in [k] \right\}.$$

Similarly, for the samples  $S$ , for true and estimated weights, define the sets of feasible assignments respectively as:

$$G_n(\ell, L) = \left\{ g : S \rightarrow \binom{[k]}{p} : \sum_{j: i \in g(x)} w_j \in [\ell, L] \forall i \in [k] \right\}$$

$$\hat{G}_n(\ell, L) = \left\{ g : S \rightarrow \binom{[k]}{p} : \sum_{j: i \in g(x)} \hat{w}_j \in [\ell, L] \forall i \in [k] \right\}.$$

**Bounding  $\alpha(S)$**  First we show that when the set  $\mathcal{X}$  is bounded in  $\mathbb{R}^q$ , then for a large enough sample  $S$  drawn from  $\mu$ , every point  $x \in \mathcal{X}$  will have a close neighbor uniformly with high probability.

**Lemma 24.** For any  $r > 0$  and any  $\epsilon > 0$ , there exists a subset  $\mathcal{Y}$  of  $\mathcal{X}$  containing at least  $1 - \epsilon$  of the probability mass of  $\mu$  such that, for any  $\delta > 0$ , if we see an iid sample  $S$  of size  $n = O(\frac{1}{\epsilon}(\frac{D\sqrt{q}}{r})^q(q \log \frac{D\sqrt{q}}{r} + \log \frac{1}{\delta}))$  drawn from  $\mu$ , then with probability at least  $1 - \delta$  we have  $\sup_{x \in \mathcal{Y}} d(x, \text{NN}_S(x)) \leq r$ .

*Proof.* Let  $C$  be the smallest cube containing the support  $\mathcal{X}$ . Since the diameter of  $\mathcal{X}$  is  $D$ , the side length of  $C$  is at most  $D$ . Let  $s = r/\sqrt{q}$  be the side-length of a cube in  $\mathbb{R}^q$  that has diameter  $r$ . Then it takes at most  $m = \lceil D/s \rceil^q$  cubes of side-length  $s$  to cover the set  $C$ . Let  $C_1, \dots, C_m$  be such a covering of  $C$ , where each  $C_i$  has side length  $s$ .

Let  $C_i$  be any cube in the cover that has probability mass at least  $\epsilon/m$  under the distribution  $\mu$ . The probability that a sample of size  $S$  drawn from  $\mu$  does not contain a sample in  $C_i$  is at most  $(1 - \epsilon/m)^n$ . Let  $I$  denote the index set of all those cubes with probability mass at least  $\epsilon/m$  under  $\mu$ . Applying the union bound over the cubes indexed by  $I$ , the probability that there exists a cube  $C_i$  with  $i \in I$  that does not contain any sample from  $S$  is at most  $m(1 - \epsilon/m)^n \leq me^{-n\epsilon/m}$ . Setting  $n = \frac{m}{\epsilon}(\ln m + \log \frac{1}{\delta}) = O(\frac{1}{\epsilon}(\frac{D\sqrt{q}}{r})^q(q \log \frac{D\sqrt{q}}{r} + \log \frac{1}{\delta}))$  results in this upper bound being  $\delta$ . For the remainder of the proof, suppose that this high probability event occurs.

Define  $\mathcal{Y} = \bigcup_{i \in I} C_i$ . Each cube from our cover not used in the construction of  $\mathcal{Y}$  has probability mass at most  $\epsilon/m$  and, since there are at most  $m$  such cubes, their total mass is at most  $\epsilon$ . It follows that  $\mathbb{P}_{x \sim \mu}(x \in \mathcal{Y}) \geq 1 - \epsilon$ . Moreover, every point  $x$  in  $\mathcal{Y}$  belongs to one of the cubes, and every cube  $C_i$  with  $i \in I$  contains at least one sample point. Since the diameter of the cubes is  $r$ , it follows that the nearest sample to  $x$  is at most  $r$  away.

Setting  $r = D\epsilon$ , we obtain one half of Lemma 10.  $\square$

For the remainder of this section, suppose that  $\mu$  is a doubling measure of dimension  $d_0$  with support  $\mathcal{X}$  and that the diameter of  $\mathcal{X}$  is  $D > 0$ . First, we shall prove general lemmas about doubling measures. They are quite standard, and are included here for the sake of completion. See, for example, [22, 21].

**Lemma 25.** For any  $x \in \mathcal{X}$  and any radius of the form  $r = 2^{-T}D$  for some  $T \in \mathbb{N}$ , we have

$$\mu(B(x, r)) \geq (r/D)^{d_0}.$$

*Proof.* Since  $\mathcal{X}$  has diameter  $D$ , for any point  $x \in \mathcal{X}$  we have that  $\mathcal{X} \subset B(x, D)$ , which implies that  $\mu(B(x, D)) = 1$ . Applying the doubling condition  $T$  times gives  $\mu(B(x, r)) = \mu(B(x, 2^{-T}D)) \geq 2^{-Td_0} = (r/D)^{d_0}$ .  $\square$

**Lemma 26.** For any radius of the form  $r = 2^{-T}D$  for some  $T \in \mathbb{N}$ , there is a covering of  $\mathcal{X}$  using balls of radius  $r$  of size no more than  $(2D/r)^{d_0}$ .

*Proof.* Consider the following greedy procedure for covering  $\mathcal{X}$  with balls of radius  $r$ : while there exists a point  $x \in \mathcal{X}$  that is not covered by our current set of balls of radius  $r$ , add the ball  $B(x, r)$  to the cover. Let  $C$  denote the set of centers for the balls in our cover. When this procedure terminates, every point in  $\mathcal{X}$  will be covered by some ball in the cover.

We now show that this procedure terminates after adding at most  $(2D/r)^{d_0}$  balls to the cover. By construction, no ball in our cover contains the center of any other, implying that the centers are at least distance  $r$  from one another. Therefore, the collection of balls  $B(x, r/2)$  for  $x \in C$  are pairwise disjoint. Lemma 25 tells us that  $\mu(B(x, r/2)) \geq (r/2D)^{d_0}$ , which gives that  $1 \geq \mu\left(\bigcup_{x \in C} B(x, r/2)\right) = \sum_{x \in C} \mu(B(x, r/2)) \geq |C|(r/2D)^{d_0}$ . Rearranging the above inequality gives  $|C| \leq (2D/r)^{d_0}$ .  $\square$

The next lemma tells us that we need a sample of size  $O((\frac{D}{r})^{d_0}(d_0 \log \frac{D}{r} + \log \frac{1}{\delta}))$  in order to ensure that there is a neighbor from the sample no more than  $r$  away from any point in the support with high probability. The second half of Lemma 10 is an easy corollary.

**Lemma 27.** For any  $r > 0$  and any  $\delta > 0$ , if we draw an iid sample  $S$  of size  $n = (\frac{2D}{r})^{d_0}(d_0 \log(\frac{4D}{r}) + \log(\frac{1}{\delta}))$ , then with probability at least  $1 - \delta$  we have  $\sup_{x \in \mathcal{X}} d(x, \text{NN}_S(x)) \leq r$

*Proof.* By Lemma 26 there is a covering of  $\mathcal{X}$  with balls of radius  $r/2$  of size  $(4D/r)^{d_0}$ . For each ball  $B$  in the cover, the probability that no sample point lands in  $B$  is  $(1 - \mu(B))^n \leq (1 - (r/2D)^{d_0})^n \leq \exp(-n(r/2D)^{d_0})$ . Let  $E$  be the event that there exists at least one ball  $B$  in our cover that does not contain one of the  $n$  sample points. Applying the union bound over the balls in the cover, we have that  $\mathbb{P}(E) \leq (4D/r)^{d_0} \exp(-n(r/2D)^{d_0})$ . Setting  $n = (2D/r)^{d_0} (d_0 \log(4D/r) + \log(1/\delta)) = O((\frac{D}{r})^{d_0} (d_0 \log \frac{D}{r} + \log \frac{1}{\delta}))$ , we have that  $\mathbb{P}(E) < \delta$ . When the bad event  $E$  does not occur, every ball in our covering contains at least one sample point. Since every point  $x \in \mathcal{X}$  belongs to at least one ball in our covering and each ball has diameter  $r$ , we have  $\sup_{x \in \mathcal{X}} d(x, \text{NN}_S(x)) \leq r$ .  $\square$

**Bounding  $\beta(S)$**  We shall now prove Lemma 11. But first, let us formally define Probabilistic Lipschitzness condition.

**Definition 4** (Probabilistic Lipschitzness). *Let  $(\mathcal{X}, d())$  be some metric space of diameter  $D$  and let  $\phi : [0, 1] \rightarrow [0, 1]$ .  $f : \mathcal{X} \rightarrow [k]$  is  $\phi$ -Lipschitz with respect to some distribution  $\mu$  over  $\mathcal{X}$ , if  $\forall \lambda \in [0, 1] : \mathbb{P}_{x \sim \mu} [\exists y : \mathbb{I}\{f(x) \neq f(y)\} \text{ and } d(x, y) \leq \lambda D] \leq \phi(\lambda)$*

*Proof of Lemma 11.* Suppose  $\mathbb{P}_{x \sim \mu}(f^*(\text{NN}_S(x)) \neq f^*(x)) \leq \epsilon$ .

Define the restriction  $f_S : S \rightarrow \binom{[k]}{p}$  of  $f \in F(\ell, L)$  to be  $f_S(x) = f(x)$  for  $x \in S$ . Firstly, we shall show that the cluster sizes of  $f_S^*$  can be bounded. Recall that the sizes of cluster  $i$  in a clustering  $f$  of  $\mathcal{X}$  and a clustering  $g$  of the sample  $S$  are respectively  $\mathbb{P}_{x \sim \mu}(i \in f(x))$  and  $\mathbb{P}_{x \sim \mu}(i \in \bar{g}(x))$ . By the triangle inequality,  $|\mathbb{P}_{x \sim \mu}(i \in \bar{f}_S^*(x)) - \mathbb{P}_{x \sim \mu}(i \in f^*(x))| \leq \mathbb{P}_{x \sim \mu}(f_S^*(x) \neq f^*(x)) = \mathbb{P}_{x \sim \mu}(f^*(\text{NN}_S(x)) \neq f^*(x))$  and this is at most  $\epsilon$ , by our assumption.

Consider  $\beta(S)$ . Since  $f^* \in F(\ell + 2\epsilon, L - 2\epsilon)$ , we have that  $f_S^* \in G_n(\ell - \epsilon, L + \epsilon)$ , we have

$$\beta(S) \leq Q(\bar{f}_S^*, c^*) - Q(f^*, c^*) = \mathbb{E}_{x \sim \mu} \left[ \sum_{i \in f^*(\text{NN}_S(x))} \|x - c(i)\| - \sum_{i' \in f^*(x)} \|x - c(i')\| \right]$$

By the triangle inequality,  $\|x - c(i)\| - \|x - c(i')\| \leq \|c(i) - c(i')\|$ . Since  $f^*(x)$  and  $f_S^*(\text{NN}_S(x))$  can differ on at most  $p$  assignments, and since any two centers are most a distance  $D$  apart, we have that  $\beta(S) \leq \mathbb{E}_{x \sim \mu}(pD \cdot \mathbb{I}\{f^*(\text{NN}_S(x)) \neq f^*(x)\}) = pD \cdot \mathbb{P}_{x \sim \mu}(f^*(\text{NN}_S(x)) \neq f^*(x)) \leq pD\epsilon$ .

All that remains is to show that  $\mathbb{P}_{x \sim \mu}(f^*(\text{NN}_S(x)) \neq f^*(x)) \leq \epsilon$  for big enough  $n$ . Lemma 28 lists the conditions when this is true.  $\square$

We require the following lemma for nearest neighbor classification, similar in spirit to that of Uner and Ben-David. Note that since  $f$  is a set of  $p$  elements, this lemma holds for multi-label nearest neighbor classification.

**Lemma 28.** *Let  $\mu$  be a measure on  $\mathbb{R}^q$  with support  $\mathcal{X}$  of diameter  $D$ . Let the labeling function,  $f$  be  $\phi$ -PL. For any accuracy parameter  $\epsilon$  and confidence parameter  $\delta$ , if we see a sample  $S$  of size at least*

- $\frac{2}{\epsilon} \left\lceil \frac{\sqrt{q}}{\phi^{-1}(\epsilon/2)} \right\rceil^q \left( q \log \left\lceil \frac{\sqrt{q}}{\phi^{-1}(\epsilon/2)} \right\rceil + \log \frac{1}{\delta} \right)$  in the general case
- $\left( \frac{2}{\phi^{-1}(\epsilon)} \right)^{d_0} \left( d_0 \log \frac{4}{\phi^{-1}(\epsilon)} + \log \frac{1}{\delta} \right)$  when  $\mu$  is a doubling measure of dimension  $d_0$

*then nearest neighbor classification generalizes well. That is, with probability at least  $1 - \delta$  over the draw of  $S$ , the error on a randomly drawn test point,  $\mathbb{P}_{x \sim \mu}(f(x) \neq f(\text{NN}_S(x))) \leq \epsilon$ .*

*Proof.* Let  $\lambda = \phi^{-1}(\epsilon)$ . We know that most of  $\mathcal{X}$  can be covered using hypercubes in the general case, as in Lemma 24 or entirely covered using balls in the case when  $\mu$  is a doubling measure, as in Lemma 26, both of diameter  $\lambda D$ . In case we have cubes in the cover, we shall use a ball of the same diameter instead. This does not change the sample complexity, since a cube is completely contained in a ball of the same diameter.

Formally, let  $\mathcal{C}$  be the covering obtained from Lemma 24 or Lemma 26, depending on whether or not the measure is a doubling measure. Define  $\mathcal{B}(x)$  to be the set of all the balls from  $\mathcal{C}$  that contain the point  $x$ . A

point will only be labeled wrongly if it falls into a ball with no point from  $S$ , or a ball that contains points of other labels. Hence,

$$\mathbb{P}_{x \sim \mu} (f(\text{NN}_S(x)) \neq f(x)) \leq \mathbb{P}_{x \sim \mu} (\forall C \in \mathcal{B}(x) : S \cap C = \emptyset) + \mathbb{P}_{x \sim \mu} (\exists y \in \bigcup_{C \in \mathcal{B}(x)} C : f(y) \neq f(x))$$

Since each ball is of diameter  $\lambda D$ , the second term is at most  $\mathbb{P}_{x \sim \mu} (\exists y \in B(x, \lambda D) : f(y) \neq f(x))$ . By the PL assumption, this is at most  $\phi(\lambda) = \epsilon$ , independent of the covering used.

For the first term, our analysis will depend on which covering we use:

- From Lemma 24, we know that all but  $1 - \epsilon$  fraction of the space is covered by the covering  $\mathcal{C}$ . When the sample is of size  $O(\frac{1}{\epsilon}(\frac{\sqrt{q}}{\lambda})^q(q \log(\frac{\sqrt{q}}{\lambda}) + \log \frac{1}{\delta}))$ , each  $C \in \mathcal{C}$  sees a sample point. For a sample this large, the first term is  $\leq \epsilon$ . Substituting  $\epsilon$  with  $\epsilon/2$  completes this part of the proof.
- When  $\mu$  is a doubling measure, we can do better. If every ball of the cover sees a sample point, the first term is necessarily zero. From the proof of Lemma 27, we know that if we draw a sample of size  $n = (2/\lambda)^{d_0}(d_0 \log(4/\lambda) + \log(1/\delta))$  samples, then every ball of the cover sees a sample point with probability at least  $1 - \delta$  over the draw of  $S$ . This completes the proof.

□